Analyzing data from single-case alternating treatments designs

Rumen Manolov[1] and Patrick Onghena[1 2]

[1] Department of Behavioural Sciences Methods, Faculty of Psychology, University of Barcelona, Spain

[2] Department of Methodology of Educational Sciences, Faculty of Psychology and Educational Sciences, KU Leuven – University of Leuven, Belgium

Running head: ATD DATA ANALYSIS

Author Note

The authors' original ideas contained in this article have not been disseminated previously in any form (e.g., conference, website). The data used in the examples were already published by other authors, as referenced. The calculations and the obtained results with the new analytical techniques applied on the published data have not been made public previously.

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain. Phone number: +34934031137. Fax: +34934021359. E-mail: rrumenov13@ub.edu

Information about Patrick Onghena, Methodology of Educational Sciences, KU Leuven,

Tiensestraat 102 - box 3762, 3000 Leuven, Belgium. Phone number: +3216325954. E-mail: patrick.onghena@ppw.kuleuven.be. At the time of this manuscript, visiting researcher at the Department of Behavioural Sciences Methods, Faculty of Psychology, University of Barcelona, Spain (patrick.onghena@ub.edu).

Running head: ATD DATA ANALYSIS

Abstract

Alternating treatments designs (ATDs) have received comparatively less attention than other single-case experimental designs in terms of data analysis, as most analytical proposals and illustrations have been made in the context of designs including phases with several consecutive measurements in the same condition. One of the specific features of ATDs is the rapid (and usually randomly determined) alternation of conditions, which requires adapting the analytical techniques. First, we review the methodologically desirable features of ATDs, as well as the characteristics of the published single-case research using an ATD, which are relevant for data analysis. Second, we review several existing options for ATD data analysis. Third, we propose two new procedures, suggested as alternatives improving some of the limitations of extant analytical techniques. Fourth, we illustrate the application of existing techniques and the new proposals in order to discuss their differences and similarities. We advocate for the use of the new proposals in ATDs, because they entail meaningful comparisons between the conditions without assumptions about the design or the data pattern. We provide R code for all computations and for the graphical representation of the comparisons involved.

*Key words***:** single-case designs; alternating treatments design; regression analysis; randomization test; trend

The field of single-case experimental designs (SCED) data analysis has seen an important growth in terms of proposals and publications, as can be seen from the summaries available in several special issues dedicated to the topic (e.g., Evans, Gast, Perdices, & Manolov, 2014; Shadish, 2014; Shadish, Rinsdskopf, & Hedges, 2008; see also Gage & Lewis, 2013). Most of the proposals for data analytical procedures are most easily (or only) applicable to what Onghena and Edgington (2005) refer to as "phase designs" in which "the whole sequence of repeated measurements is divided into treatment phases and several consecutive measurements are taken in each treatment phase" (p. 59). A reversal design, such as an ABAB design, is an example of such a phase design, as is the simple AB design, whereas a multiple baseline design is referred to as an instance of a "simultaneous replication design".

The illustrations of SCED data analysis from these special issues and from recent publications have also mainly focused on AB, ABAB, or multiple baseline designs – this is the case for nonoverlap indices (Vannest & Ninci, 2015), standardized mean difference indices (Beretvas & Chung, 2008; Shadish, Hedges, & Pustejovsky, 2014), multilevel models (Moeyaert, Ferron, Beretvas, & Van Den Noortgate, 2014), the one-level regression-based procedure (Swaminathan, Rogers, Horner, Sugai, & Smolkowski, 2014), simulation modelling analysis (Borckardt & Nash, 2014), and interrupted time-series analysis (Harrington & Velicer, 2015). The fact that phase designs are the main focus of interest is understandable, given that multiple baseline designs have been shown to be most frequently used in applied research (used in 35% of the studies according to Hammond & Gast, 2010; 54% in Shadish & Sullivan, 2011, and 69% in Smith, 2012), followed by reversal designs (used in 21% of the studies according to Hammond & Gast, 2010; 8% in Shadish & Sullivan, 2011 and 17% in Smith, 2012).

Nevertheless, for certain types of behaviors and treatments, an alternating treatments design (ATD) can be (and has been) used successfully in applied behavioral analysis and related domains: in 16% of the studies according to Hammond and Gast (2010), 8% in Shadish and Sullivan (2011), and 6% (together with simultaneous treatment designs) in Smith (2012). ATDs are characterized by a rapid and frequent alternation of conditions, which entails the absence of phases (Barlow & Hayes, 1979). Usually only one or two consecutive measurements are made under each condition, before the next switch of conditions. Given that ATDs have received less attention in terms of data analysis, our main objective is to present the analytical techniques previously proposed for ATDs as well as two new proposals that overcome some of the limitations of the existing techniques. Thus, our main questions are: *How can ATD data be analyzed?* and, more importantly, *How should ATD data be analyzed?* In order to be able to comment on the applicability and informative value of the analytical techniques and to perform a comparison among them, it is necessary (a) to present the main desirable characteristics of ATDs from a theoretical perspective and (b) to review some characteristics of the designs and the data in real applied research conducted using ATDs. In that sense, we answer the questions *When and for what purpose can the different analytical techniques be used?*, which is intended to help us choose the most appropriate analyses.

Characteristics of Alternating Treatments Designs

Main Methodologically Desirable Characteristics

In contrast to phase designs, alternation designs allow "any level of the independent variable [to] be present at each measurement occasion [and are] applicable in situations where rapid and

frequent alternation of treatments is possible" (Onghena & Edgington, 2005, p. 58). ATDs are referred to as comparative single-subject designs by Wolery, Gast, and Hammond (2010), allowing for fast comparison in readily reversible behavior. "Rapid and frequent alternation" usually means that few measurements are taken for a certain condition before changing to another condition; actually, a common restriction is a maximum of two consecutive measurements from the same condition (Heyvaert & Onghena, 2014; Kratochwill et al., 2013; Wolery, Gast, et al., 2010).

ATDs are particularly well-suited to study the effect of more than one intervention. Moreover, Wolery, Gast, et al. (2010) indicate that a control condition can also be alternated with the condition(s) of main interest in what is called the "comparison phase", instead of being only a separate initial phase. Additionally, Holcombe, Wolery, and Gast (1994) state that it is recommended to also have a final phase in which only the most effective condition is used.

The different treatments are applied in different (but contiguous) moments in time, in contrast with simultaneous treatments designs in which these interventions are available at the same time and in which the participant chooses the desired treatment (Barlow & Hayes, 1979; Barlow, Nock, & Hersen, 2009). Moreover, an ATD should be distinguished from an adapted ATD (referred to as AATD), which is designed to deal with nonreversible behaviors (e.g., when a learning process is involved). In AATDs the different conditions are applied to independent behaviors, which are supposed to be novel and of equal difficulty (Holcombe et al., 1994). The distinction is relevant for the analytical options reviewed and proposed here, given that it is common in AATD to have measurements of the different behaviors subjected to different interventions during the same measurement occasion. Thus, the number of values for each intervention is equal and there are pairs of values taking place in the same session.

The alternation of treatments is usually determined in a random way (Barlow & Hayes, 1979; Barlow et al., 2009; Kazdin, 2011; Kratochwill & Levin, 2014b). However, because the number of consecutive applications of the same condition is constrained, the corresponding randomization scheme has been called "semi-random" (Barlow et al., 2009) or "restricted" (Onghena & Edgington, 1994). Actually, the inclusion of randomization is relevant for the internal validity of the study and also for boosting the scientific credibility of the results obtained using an ATD (Edgington, 1996; Heyvaert, Wendt, Van Den Noortgate, & Onghena, 2015; Kratochwill & Levin, 2010; Tate et al., 2013; Vohra et al., 2015).

Actually, ATDs are distinguished from other designs that also include rapid and frequent alternation of treatments: the Completely Randomized Design (CRD) and the Randomized Block Design (RBD) (Edgington, 1967, 1980a; Onghena & Edgington, 1994, 2005), which perform randomization as in group-comparison experiments (see e.g., Hinkelmann & Kempthorne, 2008; Kirk, 1995), only replacing the participants by the measurement occasions as the experimental units. For example, in a CRD comparing two conditions with five measurement occasions each, there would be 252 randomization possibilities. In an RBD, the randomization possibilities are restricted to randomization within pre-specified blocks, for example by randomizing the two conditions in pairs, with one measurement occasion in the morning and one measurement occasion in the afternoon. For the comparison of two conditions with five measurement occasions, such an RBD contains only 32 randomization possibilities. This randomization scheme is identical to the randomized pair assignment that Levin, Ferron, and Kratochwill (2012) found to be associated with adequate performance of the randomization test in terms of Type I and Type II error rates, apart from representing a methodologically sound design.

CRDs are less appealing for single-case researchers because the set of randomization possibilities contains designs with undesirable properties. For the example with two conditions (A and B) and five measurement occasions each, one of the randomization possibilities for a CRD is AAAAABBBBB, precluding the necessary repeated attempts to demonstrate the intervention effect (Kratochwill et al., 2010; Onghena & Edgington, 1994). In that sense, a CRD may lead to a randomization that does not contain rapid or sufficient alternation of conditions (e.g., AAABBAABBB), or any alternation at all (e.g., BBBBBAAAAA). In some cases this lack of sufficient alternation might lead to not meeting the What Works Clearinghouse *Standards* (Kratochwill et al., 2010). Therefore, a researcher who uses a CRD may need to perform several random selections until a desirable sequence is obtained, actually performing restricted randomization.

RBDs are more popular and often the first choice in the so-called "N-of-1 randomized controlled trials" of personalized evidence-based medicine (Guyatt et al., 1990; Guyatt, Jaeschke, & McGinn, 2002; Vohra et al., 2015). However, RBDs are overly restrictive if only rapid and frequent alternation is needed. In the example above, a design such as AABBABABBA is not possible using an RBD, whereas it would be an admissible ATD. For an ATD it is only needed to define a maximum number of consecutive measurement occasions under the same condition (Onghena & Edgington, 1994).

Another option pointed out by a reviewer is to randomly choose between a sequence starting with A (e.g., ABABABAB) and a sequence starting with B (e.g., BABABABA) and then systematically alternating conditions after the first measurement occasion, but counterbalancing the two possible orders across cases. Finally, it is also possible to systematically alternate conditions, as in ABABABABAB (e.g., Morgan & Morgan, 2009). Both these options, however,

would not allow benefiting from randomization as a means of increasing internal validity (Kratochwill & Levin, 2010).

In sum, the first distinctive characteristic of ATDs is the absence of long sequences of measurements in the same condition (a minimum of three and a recommendation of five are currently endorsed for phase designs as per Kratochwill et al., 2013, and Tate et al., 2013,). This characteristic implies that levels and trends are to be estimated in a different way as compared to phase designs, as we will discuss later. Moreover, ATDs should be distinguished from RBDs, as in the former the comparison of adjacent conditions is less straightforward because there are not necessarily clear pairs to be compared. For example, in an ATD with an AABBABABBA sequence, this sequence can be split into different sets of comparisons, AABB-AB-AB-BA or AAB-BA-BA-BBA, and the analytical challenge is even greater if the number of measurement occasions for A and B is not equal. The second distinctive characteristic of ATDs is the common presence of random determination of the alternation of conditions, which makes randomization tests a natural analytical option, as discussed in a later section called “Inference”.

Design Analysis

In relation to the previously presented desirable characteristics of ATDs, Kratochwill et al. (2010, 2013) and Brossart, Vannest, Davis, and Patience (2014) stress the importance of using a design that helps ruling out threats to internal validity so that it can provide evidence for the functional relation between the behavior of interest and the manipulated variable (treatment condition).

Regarding the number of alternations, Kratochwill et al. (2010, 2013) recommend that an ATD should include five repetitions of the alternating sequence in order to meet the design

standards for providing solid evidence. Another requirement is that there are at least five data points per condition (see also Wolery, Gast, et al., 2010, who even recommend collecting data until a clear pattern is identified). The demonstration of a functional relation would, thus, require a behavioral change in the predicted direction each time that the conditions are alternated.

Regarding threats to internal validity, several threats need to be taken into account. First, "history" refers to external events occurring at the same time as the intervention and is relevant for studies gathering data longitudinally and comparing measurements before and after the change(s) in the conditions. The fact that conditions change more than once in an ATD and that the sequence of conditions is usually randomly determined, makes it less likely that external events occur always at the same moment as the change in conditions. Second, order effects (also called sequence effects) refer to the possibility that the outcomes obtained depend on the conditions being applied systematically in the same order. The random determination of the order also allows addressing this potential threat (Barlow & Hayes, 1979; Edgington, 1967, 1996), leading to many possible orders as combinations of conditions being compared (e.g., AB, AC, BA, BC, CA, CB when comparing three conditions). A systematic improvement during only one of the conditions would be a demonstration of its superiority regardless of the sequence of conditions. Third, carryover effects refer to the influence of one treatment on another subsequent treatment. This threat can be dealt with by alternating the control condition together with the intervention conditions in the comparison phase, so that it can be verified whether there are any systematic changes even in absence of an active intervention (Holcombe et al., 1994). If the behavior shows worse levels during the control condition, carryover effects are less likely. Fourth, multiple treatment interference refers to the question of whether the effect of an intervention applied in frequent alternation with another intervention would be the same if the

former is presented alone (or compared to a control condition). For dealing with this threat, it has been suggested to increase the amount of time between sessions (wash-out periods; Barlow & Hayes, 1979; Barlow et al., 2009; Kazdin, 2011). In general, internal validity threats can be tackled by including randomization and by having a large number of opportunities for a predicted effect to manifest itself or not.

Review of Alternating Treatments Designs Empirical Published Research

Aim of the review. We performed a review of published ATD studies in order to answer the following questions: (a) what are the characteristics of the design: presence or absence of randomization in determining the sequence of conditions (relevant for the performance of randomization tests; Levin et al., 2012); number of studies in which the conditions have the same number of measurement occasions (relevant for a modification of the Percentage of nonoverlapping data; PND-W); presence or absence of a baseline phase before the comparison phase (relevant for piecewise regression); (b) what are the characteristics of the data: presence or absence of overlap (relevant for visual analysis), presence or absence of linear trend (relevant for mean difference and measures of scatter based on the mean), presence or absence of nonlinear trend (relevant for mean difference and for linear regression and the possibility to apply local regression); (c) how have the data been analyzed: before reviewing and developing proposals made for analyzing ATD data, we consider that it is necessary to be acquainted with the actual practice. Additional aspects coded (which could be useful for simulation studies), but not presented here are: number of replications[1], number of conditions being compared, the average

[1] Note that in this paper, following Kratochwill et al. (2010), we use the term "repetitions" when referring to the alternation of conditions within a single ATD. We use "replication" when talking about several ATDs – across participants or across behaviors.

number of data points per condition; number of data points in each ATD; number of individual ATDs in which there were at least five measurement occasions per condition, as suggested by Kratochwill et al. (2013) and Wolery, Gast, et al. (2010).

Bibliographic search. The bibliographic search was performed in the PsycINFO database up to January 1, 2016 with the term "alternating treatments design" (in quotation marks) to be present in any field of the text. We focused on the articles published in the years 2010 to 2015, given that 2010 is the year when the What Works Clearinghouse *Standards* for SCED were published (Kratochwill et al., 2010) and it is also the year when the chapter by Wolery, Gast, et al. (2010) was published, being one of the very few recent texts explicitly discussing both the methodological and analytical possibilities for ATDs. The number of hits obtained was as follows: 23 for 2010, 26 for 2011, 29 for 2012, 28 for 2013, 27 for 2014, and 27 for 2015. We critically examined each publication to assess whether the design was actually an ATD, with the papers meeting this criterion being 8 in 2010, 7 in 2011, 10 in 2012, 7 in 2013, 6 in 2014, and 9 in 2015 (two of these nine studies were available online in 2015, but their definitive versions were published in 2016). The 47 studies reviewed represent a convenience sample in the sense that online journal articles (but not book chapters or dissertations) are included.

Operational definitions. The following operational definitions were used. For assessing whether randomization was present in the design, we read the design sections of the manuscript looking specifically for the word "random" (and its derivatives including "semi-random", e.g., Sil et al., 2013) when describing the choice of the sequence of conditions. Moreover, randomization was also judged to be present when drawing conditions from a hat (e.g., Sabielny & Cannella-Malone, 2014; Schneider et al., 2013) or when flipping a coin (e.g., Yakubova & Bouk, 2014). In contrast, when no details were provided about the order or sequence of the

conditions (e.g., Pane et al., 2015) or when only a "counterbalanced sequence" without further specification was reported (e.g., McLay et al., 2015; Mong & Mong, 2012), we considered that the design does not entail randomization.

Regarding other characteristics of the design, identifying whether the conditions had the same amount of measurements required counting the number of data points per condition. Identifying whether the ATD included an initial baseline phase consisted in inspecting the graphs for initial phases in which the behavior of interest is measured in absence of an intervention.

Regarding the characteristics of the data, overlap was defined as per Wolery, Gast, et al. (2010): only if the lines that connect the points belonging to different conditions cross, then there is overlap. Figures 1A and 1B show no overlap according to this definition, whereas Figures 1C to 1F do show overlap. In that sense, overlap was *not* defined as in the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009): whether any data point from a control condition represents an improvement over any other data point of the intervention condition, regardless of the order in the sequences.

The presence of any kind of trend, linear or nonlinear was assessed visually, instead of fitting regression straight or curve lines to assess their degree of fit. In that sense, our procedure was subjective, but we also avoided the need to compare several nonlinear models, without a clear justification for the use of any of them. Linear trend was defined as a systematic pattern of increase or decrease, meaning that, in general all measurements headed in a specific direction and did not change this direction until the end of the series or of the comparison phase. For instance, the conditions marked with an empty square on Figures 1A and 1D shows a downward linear trend, whereas 1B shows an upward linear trend. A nonlinear trend was judged to be

present when an upward trend or a downward is flattened (see the conditions marked with empty circles and empty squares in Figure 1E; see also Figure 2 of Sil et al., 2013) or when stable data initiated a trend (see the condition marked with empty squares on Figure 1C), or when there is one or more alternations between data going upwards and downwards (e.g., Figure 1F; see also Figures 2 and 3 in Yakubova and Bouck, 2014; Figure 4 in Losinski et al., 2015). Our coding reflects whether there was *any* replication of the ATD for which linear trend, nonlinear pattern, or overlap was present, as we wanted to explore what proportion of the *studies* (rather than of the individual data sets) present analytical challenges.

INSERT FIGURE 1 ABOUT HERE

Regarding the analytical indices and techniques reported in the studies reviewed, we have grouped all qualitative references to different aspects of the data (e.g., level, trend, overlap) not accompanied by numerical values into the category "visual analysis". We used the label "Mean and mean difference" for the studies comparing the level of the behavior of interest across conditions, given that some of them only mentioned the individual means, whereas others actually computed the difference between the means in the different conditions. Additionally, we separated "percentage change" from "mean difference" for those studies in which the difference is expressed in percentages (in relation to the baseline condition level) rather than in raw measures. We also grouped the different indices for dispersion (e.g., range, standard deviation) into the category "variability". In one occasion, we used the terms of the authors of the article in which it is stated that "trend analysis" was used and a quantification was provided. Finally, intervention effectiveness was assessed in some ATD studies by counting the number of sessions needed to achieve a pre-established criterion. For instance, Coleman, Cherry, Moore, Park, and Cihak (2015) implemented the following criterion: "100% accuracy for two consecutive sessions

out of three consecutive sessions in which 80% or higher responding was obtained in one condition" (p. 202). This analytical option was counted as present in our review only when the authors explicitly mentioned how many sessions were required to reach a predefined criterion.

Results. Regarding design features, the alternation of conditions is randomly determined in 25 studies (53.19%): 7 studies used an RBD, 5 incorporated a restriction about the number of consecutive implementations of the same condition, 3 mentioned counterbalancing, and 10 did not provide further information. Concerning other design features, the number of measurement occasions is the same for all conditions in all replications in 17 studies (36.17%), and an initial baseline phase is present in 25 studies (53.19%). Regarding data features, overlap was present in an ATD dataset in 44 of the studies (93.62%), linear trend was present in 41 studies (87.23%), and nonlinear trend was present in 42 studies (89.36%). These design and data features will be referenced when commenting on the different possible analyses of ATD data. Regarding the types of analysis actually applied in the published research reviewed here, Table 1 includes a summary. Specifically, visual analysis is the most commonly applied way of assessing the data; average levels and variability were present in more than half of the studies.

INSERT TABLE 1 ABOUT HERE

Alternating Treatments Designs Data Analysis

In the present section we focus on: detailing how ATD data can be analyzed, on the basis of actual practice (i.e., as found in the review presented previously) and on the basis of previously available analytical developments suggested for ATD. Additionally, after presenting all currently

existing options, we propose two new analytical techniques. For all data analysis options, we provide the following information: (a) the name of the technique and description of its application; (b) authors who have developed, adapted or proposed the technique; (c) the research question that the technique helps answering; (d) requirements about the measurement scale of the variables; (e) design requirements; (f) data patterns for which the technique is most easily interpreted; (g) possibility to compare more than two conditions; (f) possibility to compare values across studies or across replications within the same study, in case different measurement units are used; (h) relation to the information obtained in the review of published research; (i) summary of the main strengths and limitations.

Existing Analytical Techniques: Visual Analysis.

Research question and application. Visual analysis is the classical way of analyzing single case data and the most frequently applied technique, present in 75% of the studies included in our review. In fact, Barlow et al. (2009) suggest that in most cases visual analysis is expected to be sufficient for ATD data, especially if large effects (more likely to be clinically significant) are sought for. Regarding the research questions that visual analysis can help answering, Kratochwill et al. (2010) focus on its application for demonstrating evidence of a relation between an independent variable and an outcome variable.

Requirements about measurement scale of the variables, design, and data pattern. For applying the technique, the measurements should be in an ordinal, interval or ratio scale and there are no specific design requirements. Regarding the data patterns for which visual analysis is most easily applicable, Kazdin (1978) mentions that the data should not be very variable. However, as per Kratochwill et al. (2010), variability is one of the data aspects suggested to be

inspected visually, together with level, trend, immediacy of the effect, overlap, and consistency of data patterns across similar phases. Actually, in terms of applying visual analysis, Barlow et al. (2009) state that nonoverlap should be the criterion for establishing the difference between conditions, whereas levels and trends are less relevant in ATDs. In contrast, Holcombe et al. (1994) stress the importance of considering level and trend. Finally, in terms of the design features required from an ATD to apply visual analysis, and considering the usual aim of demonstrating a functional relation, we refer the reader to the "Design analysis" section presented earlier in the text.

Comparing more than two conditions in the same design and comparing values across studies. In terms of the comparison of more than two conditions represented on the same graph, the four steps of visual analysis described to the What Works Clearinghouse *Standards* (Kratochwill et al., 2010) can be applied to all pairs of conditions. In terms of comparing the results obtained across studies, the outcome of visual analysis is qualitative: according to Kratochwill et al. (2010) there is either strong, moderate or no evidence for a functional relation in an individual study. This evaluation only makes it possible to use vote-counting techniques for integrating and comparing the findings of several studies. Actually, it has been suggested that once strong or moderate evidence is obtained, statistical analysis can be applied (Kratochwill et al., 2013; Parker, Cryer, & Byrns, 2006). In the Discussion section we comment on the ways in which we consider that visual analysis can be used jointly with statistical analysis.

Summary of the main strengths and limitations. Regarding the strengths of visual analysis, the possibility to take into account all six abovementioned data features is noteworthy. However, a limitation is that applying the four steps detailed in Kratochwill et al. (2010) to ATD requires some adaptations: the first step has to refer to the measurements in the control condition

providing a clear basis for comparison, even though a baseline phase may not be present (as was the case for 46% of the studies reviewed here). For the assessment of within-phase level, trend, and variability in the second step, the use of visual aids for representing variability (e.g., standard deviation bands; Pfadt, Cohen, Sudhalter, Romanczyk, & Wheeler, 1992) or trend and a trend stability envelope (Gast & Spriggs, 2010; Manolov, Sierra, Solanas, & Botella, 2014) may lead to the ATD graphs becoming unreadable due to an excess of superimposed lines, as the measurements belonging to the same condition are usually connected to allow for comparisons. In the third step, a comparison between conditions is performed in terms of level, trend, and variability, as well as overlap, immediacy of the effect, and consistency of patterns in similar phases. In the ATD context, this would mean comparing the lines representing the different conditions. Whereas level and trend are relatively straightforward to compare, overlap is a more delicate issue. It was already mentioned that overlap can be defined as the crossing of the lines of different conditions, which is different with how overlap is defined for phase designs. Moreover, the immediacy of the effect is also not a clear criterion, as it is apparently insufficient to judge the effect of the intervention for the first alternation of conditions. Finally, the consistency in similar phases cannot be assessed in an ATD, given that there are no phases. In the fourth step, it is determined whether there are enough demonstrations of an effect at different points in time. For ATDs, at least five (rather than three) repetitions are needed, given the fast alternation of conditions. However, further clarification is required because the AABBAABBAABB example provided by Kratochwill et al. (2010, 2013) as a valid design does not apparently meet the criterion of "five repetitions of the alternating sequence" unless we understand these repetitions as "short phase transitions" (AAB-BA-AB-BA-ABB). In contrast, Kratochwill et al. (2010, 2013) also provide an example of a design with five measurement occasions for each condition

(BCBCBCBCBC) but this is only an acceptable randomized design with predetermined and fixed number of measurement occasions for only one of the random assignment possibilities. Obviously, a decision about the number of measurement occasions cannot be based on the desired randomization outcome.

As a final limitation, the performance of visual analysts has generally been assessed with phase designs (Danov & Symons, 2008) and it is unclear to what degree visual analysts would agree when inspecting more complicated graphs (e.g., Figures 1C, 1D and 1E) or graphs that do not show clear patterns (e.g., Figures 1A and 1B).

Existing Analytical Techniques: Percentage of Nonoverlapping Data.

Research question and application. One of the six data features mentioned above as critical for visual analysis has received more attention that the rest – overlap, specifically quantified via the Percentage of Nonoverlapping Data (PND; Scruggs & Mastropieri, 2013; Scruggs, Mastropieri, & Casto, 1987). In the context of ATDs, Wolery, Gast, et al. (2010) adapt it to the features of the design and advocate for its use. (In what follows we refer to their procedure as PND-W.) Despite the fact that other nonoverlap indices exist (see Parker, Vannest, & Davis, 2011, for a review), we only deal here with PND-W, as no other indices have been specifically discussed in relation to ATDs, nor used in any of the studies included in our review, where PND was used four times and PND-W once in the 47 studies.

In terms of obtaining the numerical value, the first measurement for condition A is compared to the first measurement for condition B, the second measurement for condition A is compared to the second measurement for condition B, and so forth, performing n_{min} comparisons: $n_{min} = min\{n_A, n_B\}$, where n_A and n_B are the number of measurements in each condition. The technique

quantifies the superiority of one condition as compared to another, with the quantification referring to the percentage of comparisons for which this superiority is observed and not referring to the amount of superiority in each of the comparisons (Solomon, Howard, & Stein, 2015).

Requirements about measurement scale of the variables, design, and data pattern. Regarding the measurement scale of the variables, ordinal or higher-scale data can be used. Regarding design requirements, the PND-W is best applicable when the conditions compared take place the same number of times, as in RBDs or in AATDs. For such data PND-W would allow obtaining block-by-block information about the superiority of one condition over the other. If one condition is present more than the other (according to our review only 36% of the studies included datasets in which the conditions were measured the same number of times), some data remain unused, as there is no clear indication how to proceed.

Given that PND-W does not entail estimating level or trend it does not require any specific data pattern in order for the quantification to be readily interpretable. Nevertheless, Wolery, Gast et al. (2010) recommend computing separate PND-W values for the different fractions of the data when trend is present, which makes the index in such cases ill-defined.

Comparing more than two conditions in the same design and comparing values across studies. PND-W can be applied for a comparison between all conditions pairs when more than two conditions are available, but if there is an unequal number of measurements per condition or the measurements are too distant in time (see Figure 1E) the usefulness of this index is compromised. The fact that PND-W provides a quantification in terms of a percentage means that it can be used for comparing or integrating the results across studies. Although not all

features of classical meta-analysis would be possible due to lack of knowledge regarding the sampling distribution of the index (Shadish, Hedges, et al., 2014), it is possible to use the number of data points as a weight in the meta-analysis (Shadish, Rindskopf, & Hedges, 2008).

Summary of the main strengths and limitations. The main strength of PND-W is its applicability to ordinal data and the attainment of a summary measure comparable across studies. The main limitations refer to the restricted set of conditions to which it is applicable and the lack of knowledge regarding the sampling distribution.

Existing Analytical Techniques: Mean Difference.

Research question and application. As illustrated in our review, computing means and mean differences is the most common form of quantification in ATDs, present in 72% of the studies reviewed. It is also common to accompany this quantification by some measure of dispersion (usually range): this was the case in 70% of the studies reporting means. The research question answered by using the mean difference is the magnitude of the difference between the conditions, when all values are used and no attention is paid to the sequence of the values, any existing trends, or the amount of overlap. This difference measure, expressed in the same units as the dependent variable, is usually accompanied by reporting data variability in each condition, but still without considering possible trends.

Requirements about measurement scale of the variables, design, and data pattern. Regarding measurement scale, means are meaningful for interval and ratio scale data. There are no specific design requirements for computing a mean difference. In terms of the data pattern for which the technique is most easily interpreted, a mean may summarize the data in a given condition, but it provides a poor model of the data when trends are present. Means are reasonable

as a summary measure of central tendency when trends are almost identical across conditions. In that sense, the summary provided by a mean may be missing relevant aspects of the data, such as general trends affecting the whole data series or different trends in different portions of the data. The suggestion we make here is to offer quantifications that provide information more specific than the one provided by mean. This suggestion, later formalized in our two proposals, is well-aligned with the emphasis on the importance of data variability and the fact that the average eliminates it as unimportant (Normand, 2016).

Comparing more than two conditions in the same design and comparing values across studies. Comparing more than two conditions is straightforward because each condition has its own mean. Comparability of means obtained in different studies is possible if the outcome measures are expressed in the same units or by computing a standardized mean difference (see Busk & Serlin, 1992). Two aspects need to be kept in mind when using the standardized mean difference for SCED data. First, the regular standardized mean difference for SCED data is not comparable to the standardized mean difference for group-comparison data because a measure of intra-individual variability is used as the denominator in the former and a measure of inter-individual variability is used as the denominator in the latter (Van Den Noortgate & Onghena, 2008). Smaller variability is expected when the measurements are obtained from the same individual (Beretvas & Chung, 2008). Second, the recently developed measure for standardized mean difference for SCED data, which is compatible and comparable to the standardized mean difference for group-comparison data, is not applicable to ATDs (Shadish, Hedges, et al., 2014).

Summary of the main strengths and limitations. The main strength of the mean difference is that it is straightforward to compute and interpret, allowing for comparison across studies, when standardized. The main limitation is that means provide a model that is adequate only for

stable data and a summary that would also be reasonable when the conditions compared exhibit identical linear on nonlinear patterns.

Existing Analytical Techniques: Piecewise regression.

Research question and application. Data modelling in an ATD could be considered to follow the same parametric regression-based options available for other SCEDs, so that trends can be modelled in the same step of the analysis or by first controlling for trend and then performing subsequent analysis with the residuals. However, note that none of the studies included in our review used a regression-based model.

Focusing first on simpler models, Moeyaert, Ugille, et al. (2014), and Moeyaert et al. (2015) comment on how the piecewise regression equation of Center, Skiba, and Casey (1985-1986) can be extended to be applicable to ATD. The research questions answered by piecewise regression are: (a) what is the immediate effect of introducing an intervention in the comparison phase, after a baseline phase has terminated; (b) what is the trend for each intervention; and (c) what is the difference in trends between the baseline phase and the interventions in the comparison phase. In the example provided by Moeyaert, Ugille, et al. (2014), the immediate intervention effect is estimated at two different time points after the baseline phase has finished: when conditions B starts and when condition C starts. For designs with no baseline phase (e.g., Figures 1A, 1E, and 1F), a more reasonable alternative is to perform the comparison for the last measurement time, as proposed by Shadish et al. (2013), which is also consistent with proposals for analyzing ABAB data (Olive & Franco, 2008). According to the design matrix that makes such a comparison possible, the intercept would be the fitted value for the control condition for the last

measurement occasion and it would be compared to the fitted value for the intervention condition at the same (last) moment.

Requirements about measurement scale of the variables, design, and data pattern. Regarding the requirement about the measurement scale of the variables, parametric regression is meaningful for interval and ratio scale data. In terms of design requirements, the adaptation of piecewise regression, as described by Moeyaert et al. (2014), requires an initial baseline phase.

Regarding the data pattern for which the technique is most easily interpreted, the illustrations by Moeyaert, Ugille, et al. (2014) and Moeyaert et al. (2015) deal with linear trends, but modelling could also be performed via generalized additive models in such a way that different trends and different data patterns are taken into account (Shadish, Zuur, & Sullivan, 2014), with the use of a Poisson model specifically suggested for count data (Shadish, Kyse, & Rindskopf, 2013). In terms of most suitable data patterns, in some cases the interpretation of results expressed as an immediate change in intercept at the beginning of the comparison phase and a difference in slopes can be challenging: (a) when the intercept is higher (desirable) for one of the conditions but the trend is decreasing (see Figure 1C); and (b) when the lines connecting the measurements belonging to different conditions cross (Figure 1D). Moreover, if the last measurement occasion is chosen for comparing levels, the difference observed for the data may not represent the complete data patterns in an adequate way (e.g., Figures 1C and 1D).

Comparing more than two conditions in the same design and comparing values across studies. Regarding application for comparing more than two conditions, it is possible to compute the change in intercept (whichever point it is defined for) for all comparisons between pairs of conditions; it is also possible to perform pairwise comparisons between estimated trends. For

comparing values obtained in different studies using different measurement units for the dependent variable, Van Den Noortgate and Onghena (2008) proposed standardizing the effects by dividing them by square root of the mean square error[2], that is dividing by $\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-1)}$, where y_i are the actual observations and $\hat{y}_i$ are the predicted values. These standardized values can be compared and used in meta-analysis via multilevel models.

Summary of the main strengths and limitations. The main strengths of piecewise regression are modeling flexibility and possibility to compare and integrate results across studies. The main limitations are the relative complexity of the models (i.e., the definition of the design matrix), the applicability limited to data patterns for which a comparison of intercepts in one measurement occasion (which may not be the same across studies) is meaningful and not misleading, and due to the lack of single overall quantification, the interpretation may not be straightforward when differences in intercepts and slope are in opposite directions.

Existing Analytical Techniques: Local Regression.

Research question and application. The use of local regression (LOESS) is motivated by the fact that observed trends may not be sufficiently well represented by a straight line or a second-order (quadratic) polynomial model (see Figure 1F and illustration 6 from the online supplementary material). Specifically, Solmi, Onghena, Salmaso, and Bulté (2014a) propose to use nonparametric smoothers for fitting curves to the measurements in each condition and comparing those curves; LOESS allows surpassing the need to specify a priori the type of relation between time and measurements (Jacoby, 2000). LOESS requires choosing a linear or a

[2] Obtained in R via the command sqrt(sum((residuals(reg)^2))/df.residual(reg)) on a previously saved object "reg" including the results from the piecewise regression analysis, or for the polynomial or LOESS regression analysis for the data in each condition separately.

quadratic model as the basis for each local regression and to deciding the fraction of the data to use (see R. A. Cohen, n.d.; Hurvich, Simonoff, & Tsai, 1998) in each local regression via the smoothing parameter. Regarding the research question that the technique helps answering, LOESS quantifies the difference between the conditions represented by straight or curved lines that capture some of the observed variability in the data, but not potential outliers, or all of the variability in the data, if the model provides a perfect fit to the measurements.

Requirements about measurement scale of the variables, design, and data pattern. Regarding measurement scale, given that each local regression could be linear or quadratic, it is necessary that the data are measured in an interval or a ratio scale. In terms of design requirements, it is not necessary that there is the same number of measurements per condition. However, the evidence provided by Solmi et al. (2014a) indicating appropriate performance in terms of Type I error and statistical power refers to the series lengths (30-100) that are longer than the ones observed in the current review of ATD research (values ranging from 2.83 to 17.5 average measurements per condition in a study, with an overall mean of 6.60 and median of 5.38). Thus, it is not clear whether the procedure will perform well with typical ATD data. This uncertainty as well as the fact that the proposal for applying LOESS to ATD data is recent could be among the reasons for not finding any applications of LOESS among the 47 studies reviewed here. In terms of the data pattern for which the technique is most easily interpreted, a theoretically-supported model for the relation between time and measurements is not necessary, nor is it required to assume a specific form of this relation before the analysis.

Comparing more than two conditions in the same design and comparing values across studies. Regarding the application of LOESS for comparing more than two conditions, a separate curve is fitted to the data from each condition before comparing them in a pairwise

fashion. Regarding the possibility to compare effects across studies, for piecewise regression the differences in intercept and slope can be standardized on the basis of the variability of the residuals. For LOESS, however, there are as many sets of residuals as conditions. In this case, it is possible to standardize each value using the same procedure mentioned before (Van Den Noortgate & Onghena, 2008). Running the analyses again with the standardized data will lead to mean differences between predicted values that are comparable across ATD replications. For the combination of results obtained via the nonparametric smoother, see Solmi, Onghena, Salmaso, and Bulté (2014b).

Summary of the main strengths and limitations. The main strength of LOESS is the possibility to model the data in each condition without assuming any specific data pattern a priori and without requiring the same number of measurement occasions per condition. The main limitations of the procedure are the subjective and potentially not replicable decisions made for choosing among different possible models.

New Proposal: ADISO

Rationale, research question and formal representation. In the present paper we propose a new analytical procedure consisting in comparing adjacent conditions, each of which usually contains one or two measurement occasions, and in obtaining the weighted average of the differences observed in all comparisons. We refer to this proposal as "average difference between successive observations" (ADISO) and we offer R code for its computation and graphical representation in the online supplementary material, apart from a user-friendly website that also incorporates ADISO (http://manolov.shinyapps.io/ATDesign).

The research question answered refers to the average difference between the conditions, when the comparisons are performed on the basis of actually obtained measurements and include only measurements of adjacent conditions. As an example, consider that data from Figure 1F, representing an AABBAABBABABBABB design. One possible set of comparisons between adjacent conditions would be AABB-AABB-AB-ABB-ABB. In such a situation, the mean of the first AA pair can be compared to the mean of the second BB pair and so forth until the last comparison comprising the last A measurement with the mean of the last two B measurements (see the graphical representation on Figure 2A). There would be five differences (8.15, 6.75, 6.60, 24.95, and 9.80) and the value of ADISO is their weighted average (11.07), with weights representing the number of measurements involved in the comparison (4, 4, 2, 3, and 3 in the AABB-AABB-AB-ABB-ABB partition). The differences for each comparison show how the distance between the conditions varies as the data series progresses; a piece of information not provided by the mean difference or by PND-W.

Another way of conceptualizing ADISO is as a difference between the weighted averages of the measurements belonging to the conditions being compared, with the weights representing the importance of the value in the comparison. If positive signs are arbitrarily assigned to the values of the A condition and negative signs to the values of the B condition, without affecting the final result, the weights for the AABB-AABB-AB-ABB-ABB partition represented in Figure 2A would be, as follows: the first comparison AABB entails 4 values, there are two A condition values each of which is assigned a weight of $4/2 = 2$ and two B condition values, each of which is assigned a weight of $-(4/2) = -2$; the second comparison is identical to the first one; the third comparison is AB, entailing 2 values, and the weight for the A condition value is $2/1 = 2$ and for the B condition value is $-(2/1) = -2$; the fourth and fifth comparisons are

identical, ABB, including 3 values, and the weight for the A condition value is $3/1 = 1$ and for each of the B condition value is $-(3/2) = -1.5$.

For the AAB-BA-ABB-AB-ABB-ABB partition represented in Figure 2B the weights would be as follows: the first comparison AAB entails 3 values, there are two A condition values each of which is assigned a weight of $3/2 = 1.5$ and the B condition value is assigned a weight of $-(3/1) = -3$; the second and the fourth comparisons both entail two values, one per condition, the weight for the A condition value is $2/1 = 2$ and for the B condition value is $-(2/1) = -2$; the third, fifth, and sixth comparisons are all ABB, including 3 values, and the weight for the A condition value is $3/1 = 3$ and for each of the B condition value is $-(3/2) = -1.5$.

In general:

$$ADISO = \frac{\sum_{i=1}^{n_A} y_{Ai} \times w_{Ai}}{\sum_{i=1}^{n_A} w_{Ai}} - \frac{\sum_{j=1}^{n_B} y_{Bj} \times w_{Bj}}{\sum_{j=1}^{n_B} w_{Bj}},$$

$$where \begin{cases} w_{Ai} = \left(n_{comp(A)} + n_{comp(B)}\right)/n_{comp(A)} \\ w_{Bj} = -\left(n_{comp(A)} + n_{comp(B)}\right)/n_{comp(B)} \end{cases},$$

where n_A and n_B represent respectively the number of measurements in conditions A and B, $n_{comp(A)}$ and $n_{comp(B)}$ represent respectively the number of measurements from conditions A and B that that are used in the comparison in which the values y_{Ai} and y_{Bj} participate, and w_{Ai} and w_{Bj} represent the weights assigned to each value from condition A (y_{Ai}) and each value from condition B (y_{Bj}). Considering this expression, the simple mean difference refers to the case in which all w_{Ai} are equal among themselves and all w_{Bi} are equal among themselves, which makes the weight of each A value equal to $1/n_A$ and the weight of each B value equal to

$1/n_B$. In that sense, ADISO assigns a greater weight to a value that is critical (i.e., the only one from its condition) in the context of the comparison in which it is involved, while also taking into account the number of values used in this specific comparison. In contrast, the simple mean difference assigns greater weights to values from conditions with fewer measurements in general. This is an illustration of the more specific emphasis that ADISO has on the comparisons actually being performed.

Possibility for an ordinal comparison. So far ADISO has been presented as a way, alternative to the simple mean difference, for quantifying the distance between two conditions, focusing on the comparisons between measurements pertaining to adjacent conditions rather than using overall averages. Nevertheless, ADISO can also be considered as an alternative to PND-W, given that for each comparison, it is possible to only count whether the A or the B condition is superior in ordinal terms, without computing the difference in the measurement units of the dependent variable. Thus, the overall ordinal quantification, ADISO-O, would be the percentage of comparisons for which B is superior to A; only focusing on adjacent comparisons (rather than comparing values that are in the same position in the sequence of values from their own condition, like PND-W) and being applicable also to data for which $n_A \neq n_B$ (unlike PND-W). In that sense, ADISO-O for condition B being superior to condition A could be formally defined as

$$\frac{\#(\bar{y}_A < \bar{y}_B)}{c} \times 100\%,$$

where c is the number of comparisons performed, $\bar{y}_A$ is the average of A condition values that are used in a given comparison, $\bar{y}_B$ is the average of B condition values that are used in the same comparison, and # represents counting the number of comparisons for which the condition is met.

For the AABB-AABB-AB-ABB-ABB partition represented in Figure 2A, the B values are lower than the A values in all five comparisons leading to 100% superiority. In contrast, in the AAB-BA-ABB-AB-ABB-ABB partition represented on Figure 2B, the B condition has lower values for 5 of the 6 comparisons, leading to 83.33% superiority.

INSERT FIGURE 2 ABOUT HERE

Partitioning the data sequence. As the previous examples have shown, there is not always a single way of defining which adjacent comparisons to perform. In that sense, there are four options for choosing how to segment the sequence of data. First, it is possible to choose a segmentation that is meaningful according to substantive criteria, as when an RBD is used (blocks representing the natural segmentation points) or using the information about when the measurements were taken (e.g., comparing data points from different conditions obtained on the same day). We recommend this criterion as the first option as it is based on the design actually used or, when the day of measurement is used as a basis, it allows for more control of a potentially extraneous variable such as whether the individual had a good or a bad day.

Second, it is possible to perform all comparisons of conditions being present in the same order: for the Figure 1F data, this would lead to five comparisons AABB-AABB-AB-ABB-ABB (Figure 2A); for the Figure 1E data comparing empty squares (C) and filled triangles (B), this would lead to CBBB-CCBB-C. The interpretative advantage is that all quantifications refer to switching from condition A to condition B. However, there are two issues: (a) the rapid (and frequently randomly determined) alternation of conditions is supposed to make the order of the

conditions irrelevant and to counter sequence effects; and (b) as in the example for the Figure 1E data, this approach might lead to some unused measurements.

Third, it is possible to choose a segmentation that leads to more comparisons being performed: for the Figure 1F data, the segmentation AAB-BA-ABB-AB-ABB-ABB (shown on Figure 2B) leads to six comparisons and, for the Figure 1E data, the segmentation CBB-BC-CB-BC leads to four comparisons. As illustrated, the advantage of such an approach that it could favor meeting the What Works Clearinghouse *Standards* of five repetitions of the alternation or meeting the standards with reservations – four repetitions (Kratochwill et al., 2010).

Fourth, it is possible to avoid making a decision, by applying ADISO for all possible segmentations (or only for those meeting design standards, if one or several do) and exploring the extent to which the value of ADISO differs according to the quantification. In that sense, a sensitivity analysis would be performed, as suggested for multilevel models (Ferron et al., 2008). If results do not differ greatly, a reasonable approach would be to compute the average value of ADISO across all segmentations. In contrast, if the results are very different, reporting all ADISO values and tentatively interpreting them is the only option. We consider that this latter approach is the second best option, in case the features of the design cannot be used for determining the partition of the sequence.

Requirements about measurement scale of the variables, design, and data pattern. Given that ADISO entails computing means an interval or ratio scale is required for the measurements of the dependent variable. However, ADISO-O can be computed for ordinal data as well. In terms of design, there are no specific requirements about series length or the number of measurements per condition, or the necessity of an initial baseline phase. For AATD in which

there are measurements for each intervention available for the same measurement occasions, it is not necessary to choose a partition of the sequence because the natural comparison is between values from the same session. Regarding the data pattern for which ADISO is most suitable, the fact that no specific relation between time and the measurements is assumed, ADISO is applicable to stable data, as well as to data exhibiting linear or nonlinear trends.

Application of ADISO when comparing more than two conditions. In case more than two conditions are being alternated, there are two possible approaches. The first approach is to compare each occurrence of the preceding condition with each subsequent occurrence of another condition. For instance, for the data in Figure 1A (ABCBCABACCBA), it is possible to compare AB three times, AC or CA four times, and BC or CB four times. The second approach is to perform only the comparisons between contiguous measurements (i.e., AB-CB-CA-BA-CCB-A for the data in Figure 1A). The advantage of this second approach is that it involves comparisons that are better aligned with ADISO's logic of comparing only adjacent conditions, but the drawback is that there are fewer comparisons and, as in the example, it is possible that some measurement remains unused. We advocate for the second approach in order to avoid comparing conditions separated by other conditions in the sequence. According to our review of published research using ATD, it is most common to compare two or three conditions (resp. 47% and 49% of the studies); in 4% of the studies four conditions were compared. If pairwise comparisons are performed, this would lead to six possible comparisons, which could be problematic if these comparisons were accompanied by multiple unadjusted statistical tests (i.e., randomization tests) because it would increase the probability of obtaining a statistically significant result by chance. In such cases, a Holm-Bonferroni or a Dunn-Sidák adjustment could be used to control for the

family-wise Type I error rate (Edgington & Onghena, 2007; Westfall & Young, 1993). However, the descriptive use of ADISO would not be compromised.

Comparing values across studies. The main outcome of ADISO is expressed in the same measurement units as the dependent variable, which limits the comparison across studies using different operative definitions of the same constructs. The standardization we propose here for making values comparable consists in dividing the ADISO value by the standard deviation of the differences computed for each comparison, which are averaged to obtain the ADISO value itself. For instance, for the Figure 2A data, the comparisons led to the following differences 8.15, 6.75, 6.60, 24.95, and 9.80, whose standard deviation is 6.95, which would lead to a standardized ADISO of 1.59. In comparison, the standardized mean difference using $\sqrt{\left((n_A-1)s_A^2+(n_B-1)s_B^2\right)/(n_A+n_B-2)} \approx 9.80$ in the denominator yields 1.15. ADISO-O is expressed as a percentage and thus is comparable across studies. In terms of meta-analysis, the sampling distribution of ADISO and ADISO-O have not been derived, and thus classical meta-analysis is not possible, but weighted averages can be obtained using the series length as a weight (as was the case for PND-W).

Summary of the main strengths and limitations of ADISO. The main advantages of ADISO are the use of meaningful comparisons between adjacent values, the lack of design and data pattern requirements, and the possibility of quantifying distance for interval or ratio scale variables and quantifying superiority for ordinal variables. Thus, ADISO is more generally applicable than PND-W. The main limitation of ADISO is the choice of how to segment the sequence, although some recommendations were provided, and the unknown standard error of the values. The segmentation problem also entails a practical issue that limits its use as a test

statistic in a randomization test, described later in the "Inference" section, as R code has still not been developed for performing all possible segmentations for the actual sequence of conditions and for all conditions that could have been obtained at random, according to the randomization scheme.

New Proposal: ALIV

Rationale, research question and formal representation. In the present paper we propose a second novel analytical procedure consisting of numerically comparing the values that are represented by the lines used to connect the measurements belonging to different conditions. These values include both actually obtained values (i.e., the dots in a graph) and linearly interpolated values (i.e., the dots that could be placed on the line, representing possible values in case the condition had taken place during a measurement occasion in which the other condition was present). We refer to this procedure as ALIV (actual and linearly interpolated values), with the main quantification being the differences for the $n = n_A + n_B$ measurement occasions. An illustration is provided on Figure 3, created with the R code we offer in the online supplementary material.

INSERT FIGURE 3 ABOUT HERE

Whereas ADISO and ADISO-O were proposed as alternatives to the mean difference and PND-W, ALIV is proposed as an alternative to the mean difference and to linear, quadratic or LOESS regression models. Specifically, in contrast with the mean and similarly to ADISO, ALIV provides quantifications that illustrate how the difference between conditions varies across

different portions of the data series. Additionally, in comparison to LOESS, ALIV allows avoiding the subjective decision of how well the LOESS model should fit the data, a decision that cannot be aided statistically via an *F* test or a Bayesian Information Criterion. Actually, in case a perfect fit is desired from LOESS, the result would be identical to ALIV. In that sense, we consider that a simple linear interpolation would be more parsimonious and in certain cases equivalent to a LOESS model (compare illustrations 1, 3, 4, 5 to illustration 2 and 6 in the online supplementary material). Therefore, the research question answered is: how much is the difference between the lines connecting the points belonging to different conditions. Another way of conceptualizing the research question is: what would be the average difference between conditions, if the actually obtained measurements in one condition are compared to counterfactual values from the other condition, estimated via linear interpolation.

Note that ALIV is different from ADISO, given that in ALIV the comparisons are performed for the same measurement occasions, comparing actual with interpolated values in an alternating way. In contrast, for ADISO adjacent actually obtained values are being compared. Moreover, ALIV does not include the first and the last measurement occasions in the comparison because the data points for these measurement occasions cannot be interpolated for the condition that is not taking place. However, the first and the last measurements are used in the interpolation of the contiguous values for the condition(s) that take place during these measurement occasions. For instance, in Figure 1C, condition A takes place on sessions 1 and 3 and, thus, the value for session 1 is not used in the comparison, but it is used (together with the measurement for session 3) to interpolate the A condition value for session 2.

Formally, ALIV can be understood as a difference between the weighted averages of the measurements belonging to the conditions being compared. Specifically, the weights reflect

whether the specific measurement is included in the comparison ($m = 1$ if it is not outside of the fraction of measurement occasions used in the comparisons; see Figure 3) and the number k of values that are interpolated using the specific measurement. Formally:

$$ALIV = \frac{\sum_{i=1}^{n_A} y_{Ai} \times w_{Ai}}{\sum_{i=1}^{n_A} w_{Ai}} - \frac{\sum_{j=1}^{n_B} y_{Bj} \times w_{Bj}}{\sum_{j=1}^{n_B} w_{Bj}},$$

$$\text{where } \begin{cases} w_{Ai} = m + 0.5\,k \\ w_{Bj} = -(m + 0.5\,k) \end{cases}$$

$$\text{and } \begin{cases} m = 0 \;\; for\ values\ before\ or\ after\ all\ values\ from\ the\ other\ condition \\ m = 1\ otherwise \end{cases},$$

where n is the total number of measurement occasions in the comparison phase ($n = n_A + n_B$), k is the number of interpolated values in the determination of which the specific actually obtained value (y_{Ai} or y_{Bj}) participates. In the previous expressions, we have arbitrarily assigned negative signs to the B condition values and positive signs are assigned to the A condition values, but this does not change the final results.

For instance, for the data from Figure 3, the weight of the first value is equal to $m = 0$ plus 0.5 times $k = 0$, $w_{A1} = 0$, as it is outside of the fraction used for the comparisons and not used for linearly interpolating any value; the weight of the second value is $m = 0$ plus 0.5 times $k = 2$, $w_{A2} = 2$, as it is outside the fraction used for the comparisons, but it is used for linearly interpolating the following two A condition values. Analogously, all n weights are obtained: 0, 1, −1, −2, 2, 2, −2, −1.5, 2.5, −2, 2.5, −1.5, −1.5, 2, −0.5, and 0.

In comparison to the simple mean difference, ALIV assigns greater weight to measurements farther away from other measurements in the same condition. Such measurements are considered

more important as they serve as pivotal points for assessing the performance in a given condition at that point in time, because they are the only piece of information available (see the first filled triangle in Figures 1C and 1D). A potential drawback of such greater weights assigned to values isolated from other values of the same condition would be assigning a greater weight to isolated values that could be outliers (e.g., the penultimate empty circle in Figure 1F).

Requirements about measurement scale of the variables, design, and data pattern. Given that ALIV entails computing means, variables in an interval or ratio scale are required. In terms of design, there are no specific requirements about series length or the number of measurement occasions per phase. As ALIV entails not using the first and/or the last values of a series, in case these values are followed and preceded, respectively, by measurements from the same condition, more data would be lost for cases such as the ones depicted on Figure 1D and Figure 1F, in comparison, for instance to sequences such as the ones from Figures 1B and 1C. Regarding the applicability to AATD in which there are measurements for each intervention available for the same measurement occasions, it is not necessary to interpolate values and, thus, no data points remain unused. Actually, the ALIV value would be equal to ADISO and to the simple mean difference. Regarding the most easily interpretable data patterns, the application of ALIV does not assume stable data, or linear or any specific nonlinear trend.

Application of ALIV when comparing more than two conditions. In case an ATD compares more than two conditions (e.g., Figures 1A and 1E), the ALIV can be applied by comparing all pairs of conditions (e.g., AB, AC, and BC, when there are three conditions). However, researchers should be cautious when interpolating several values only on the basis of the straight line that connects only two actually obtained measurements (see the distance between the first and second empty square in Figure 1D).

Comparing values across studies. As was the case for ADISO, the main outcome is expressed in the same measurement units as the dependent variable. For making comparisons across studies possible, we propose standardizing by dividing the ALIV value by the standard deviation of the differences computed for each comparison, which are averaged to obtain the ALIV value itself. For instance, for the Figure 3 data, the comparisons led to the following differences 12.2, −2.5, −3.4, −6.9, −5.7, 5.7, 5.1, 21.5, 29.9, 14.4, 8.9, and 4.9, whose standard deviation is 10.72, which would lead to a standardized ALIV of 0.65.

Summary of the main strengths and limitations. The main strengths of ALIV are: (a) it enables a quantitative analysis that mimics the visual inspection of the data, based on the lines connecting points from the same condition, representing the comparison between observed and projected patterns across all data (Kratochwill et al., 2010); (b) the only assumption is that the neighboring values are the best option for estimating the measurements that could have been obtained between them; (c) no model has to be specified a priori; (d) no decision is required regarding the measurement occasion for which a comparison in intercept can be performed (unlike piecewise regression); and (e) the application does not require subjective decisions as would be the case when using LOESS. The main limitation of ALIV is not using the first and last values in the sequence for the comparison, although such values can be used for interpolating other values.

Inference

In the previous sections we reviewed several quantifications of the magnitude of difference between conditions and proposed two new such quantifications. All these quantifications could

be labelled “effect size” measures to be used as descriptive measures for a specific data set, similar to the use of an arithmetic average or a median, to describe the central tendency of a data set without invoking any additional assumptions. Actually, in relation to the term “effect size”, several definitions have been provided regarding what constitutes an effect size index (e.g., strength of relationship between an independent and a dependent variable, the magnitude of the impact of a treatment on an outcome measure). After a thorough review of such definitions, Kelley and Preacher (2012) define an effect size as “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest” (p. 140) and the effect size index is the equation that defines the dimension of interest in an operational way. Following the discussion by Kelley and Preacher (2013) and Carter (2013), it is crucial that the effect size index quantifies a specific dimension, despite the fact that it may not be sensitive to other kinds of effect. An effect size can be unstandardized, when a common and meaningful metric is used across studies (e.g., weight loss in kilograms) or standardized when the response variable is not measured in the same measurement units (Lipsey & Wilson, 2001), with nonoverlap indices not requiring standardizing as they already entail a common metric.

When the focus is put on the actually obtained data, these constitute the population of interest. In recent publications about these effect size measures, their statistical properties as effect size “estimates” have been discussed (Kratochwill et al., 2010, 2013; Shadish, Rindskopf, & Hedges, 2008). In an estimation context, the measures are used to obtain information about an unobserved value, conceptualized as a population parameter. In this case, it is crucial to know which population is at stake (a certain population of similar cases or the population of past, present, and future outcomes of a particular case) and which random processes are assumed or involved to quantify the uncertainty surrounding the estimate. Usually, an assumption of random sampling is

needed to firmly ground the statistical inferences from the sample to the population (Edgington & Onghena, 2007; Kempthorne, 1979).

Because true random samples are rare in applied research and seem difficult to reconcile with single-case research, we could take another approach, and focus on the functional relation between the manipulated independent variable (X) and the outcome variable (Y). The inferential question in this approach is whether the relation is causal. In other words, we are considering "causal inference" instead of "sample-to-population inference". In causal inference we derive a probabilistic statement, conditional on the null hypothesis that there is no causal relation between X and the outcome variable Y. This statement is made possible by the joint use of randomization in the design and a randomization test for data analysis. A tentative (i.e., a probabilistic and cautious) causal inference is possible thanks to the confluence of: (a) using an experimental design that controls for as many known confounding factors as possible, (b) incorporating randomization in this design to control for known and unknown confounding factors that are time-related, (c) using a test statistic that is sensitive to the predicted effect, and (d) using a randomization test for quantifying the probability of obtaining a difference as large as the one obtained only by chance (Dugard, File, & Todman, 2012; Edgington & Onghena, 2007; Ferron & Levin, 2014; Kratochwill & Levin, 2010).

Suppose that X has only two levels (Treatment A and Treatment B), then the causal effect of X on outcome Y in an SCED can be defined as the difference in Y between Treatment A and Treatment B *at any given measurement occasion*. However, just as it is impossible in a between-subjects group comparison study to observe a subject in the experimental and the control condition simultaneously, it is equally impossible to have a measurement occasion in an SCED in which *both* Treatment A and Treatment B are implemented and can be compared

independently. In technical terms: only one of the Y scores is observed; the other one is missing. This missing Y score is called the counterfactual or potential outcome. Now the interesting part is that we know this counterfactual outcome in a randomized design if the null hypothesis is true: the Y score would just have been the same if another assignment was selected. Consequently, we can validly use a randomization test in a randomized design to derive a *p*-value, given that the null hypothesis is true (Edgington & Onghena, 2007; Holland, 1986; Rubin, 1974, 2005).

Such a causal inference has limited but clear ambitions. Its focus is on internal validity and statistical-conclusion validity. In the absence of random sampling, external validity cannot be based on statistical inference. For external validity one must rely on theoretical argument (Eisenhardt, 1989; Yin, 2014), comparison of the context and circumstances of the experiment and abduction (Evers & Wu, 2006), replication, falsification, and corroboration (Barlow et al., 2009; Flyvbjerg, 2006), or systematically ruling out the major threats to external validity, of which the "Interaction of the Causal Relationship with Units" is probably most relevant for single-case research (Shadish, Cook, & Campbell, 2002).

Taking into account the way in which the treatment conditions are assigned to the measurement occasions, such a causal inference accompanied by a randomization test is a natural analytical option for ATD data (Edgington, 1980b; Onghena & Edgington, 1994). Even with short data series, the randomization test remains valid and with an ATD the number of possible and acceptable random assignments is large enough to ensure the possibility to obtain statistically significant results (Onghena & Edgington, 2005). Actually, the validity of the randomization test is based on the requirement of random assignment in the design (e.g., random choice of the points of change in phase in an ABAB design, random choice of the sequence of conditions in an ATD) prior to collecting the data. The randomizations performed after the data

are gathered, needed for obtaining the reference distribution to which the test statistic is compared, have to correspond to the random assignment scheme actually used (Edgington, 1980b). For example, in the Sil et al. (2013) study, the conditions were determined "semi-randomly" (p. 332), meaning that a restriction of a maximum of two consecutive administrations of the same condition was introduced when randomly determining the order of conditions. Considering the researchers' decision that $n_A = n_B = 5$, this leads to 84 possible sequences, for instance, not including randomizations such as AAABBBABABB, which would stem from a completely randomized design, but which could not have been obtained by the random assignment procedure followed by the researchers.

An assumption necessary for performing the randomizations after the data are collected is the exchangeability of the data (Hayes, 1996). In a randomized ATD this exchangeability is guaranteed by the actual random assignment procedure. If the null hypothesis is true, the same measurements would have been obtained, whatever treatment condition was applied at each measurement occasion. Serial dependencies that are common in time series data do not pose a problem for randomization tests because these serial dependencies are constant for all possible random assignments if the null hypothesis is true.

The main output of a randomization test is a *p*-value for the null hypothesis that there is no causal effect (i.e., no difference between the conditions). The effect itself is quantified using a test statistic. Regarding the choice of a test statistic, randomization tests are flexible enough to allow choosing it according to the aims of the researcher and the effect expected. For instance, it is possible to use a nonoverlap index, a difference in means, or a difference in trends (Heyvaert & Onghena, 2014). In the current paper, we argue for using ADISO or ALIV to test for a difference in average level between the conditions in an ATD. Furthermore, if an assumption can

be made about the form of the causal effect (e.g., a constant additive effect), then also a confidence interval around the effect size can be constructed based on randomization test inversion (Heyvaert & Onghena, 2014; Michiels, Heyvaert, Meulders, & Onghena, 2016).

Finally, the *p*-values yielded by randomization tests can be combined using several different approaches (Rosenthal, 1978). A practical approach included in the SCDA plug-in for R (Bulté & Onghena, 2012) is Edgington's (1972) additive method. Technically, the combined *p*-value that results from this additive method represents the probability, under the null hypothesis of no difference between the conditions, of getting such a small sum of probabilities as the sum actually obtained. This probability can be used to assess whether it is likely that the differences (across replications) observed between the conditions is only due to chance variations.

In sum, the main strengths of randomization tests are: (a) the possibility of valid inference about causality; (b) the flexibility in choosing the test statistic according to the effect of interest or what is expected on the basis of previous knowledge; and (c) the possibility to integrate the results of several studies via combining *p* values. The main limitations of randomization tests are: (a) the requirement of random assignment in the design, but our review shows that the alternation of conditions in the ATDs was decided at random in more than 50% of the studies; and (b) the fact that they are computer-intensive, which could be the reason for their underuse observed in our review of ATD studies, but with present-day availability of fast computers and user-friendly software, this problem has been largely overcome (see e.g., Bulté & Onghena, 2012, 2013; Levin, Evmenova, & Gafurov, 2014).

Applying the Analytical Options

In Table 2 we provide a summary (and, thus, a simplified representation) of the main features of the analytical procedures discussed and proposed in the present text and we have also applied the analytical techniques to the ATD datasets from Figure 1 and we provide the results of the analyses and graphical representations in color in the online supplementary material. We also mention the randomization scheme used for determining the alternation of conditions and provide the results for randomization tests for some of the examples in which different schemes are used.

INSERT TABLE 2 ABOUT HERE

In case the data are relatively stable and show no overlap (e.g., the Andersen, Daly III, and Young, 2013, data for Terrance represented in Figure 1A; illustration 1 from the online supplementary material), all techniques are readily applicable and lead to very similar results. Therefore, simple procedures such as the mean difference can be computed, although the fact that trends are not identical suggests that piecewise regression, providing a good fit to these data, can offer more nuanced information about trends and difference in intercept. When the data pattern is straightforward, the choice of an analytical technique is not critical. In the Andersen et al. (2013) study the random assignment procedure can be conceptualized as an RBD, given that "all conditions were administered in random order before they were readministered a second, third, and then a fourth time, each time in random order" (p. 407). For pairwise comparisons of conditions, this leads to only $2^4 = 16$ possible randomizations, as there are four occasions for randomly choosing between two possible orders (AB or BA). With 16 randomization it is impossible to attain $p \leq .05$ because the minimum p-value is $1/16 = 0.0625$.

In case the data show no overlap, but there are markedly different trends for the different conditions (e.g., the Coleman et al., 2015, data for Alice represented on Figure 1B; illustration 2 from the online supplementary material), the results between the procedures quantifying mean differences are still very similar. Nevertheless, simple models like the mean and first and second-order polynomial regression may not be appropriate when data are so variable. In contrast, LOESS, using a fraction of 60% and a linear model for each regression, provides better fit and its results are practically identical to the ones obtained by ALIV, which does not require making any arbitrary decisions. Finally, in this case ADISO provides more conservative results than the mean difference based on actual data and the mean differencing arising from regression analysis. In terms of the percentage of comparisons for which one condition is superior to the other, the results of ADISO-O and PND-W are similar. In terms of the application of a randomization test, the procedure followed in the Coleman et al. (2015) is equivalent to an RBD, given that, for each participant, an online randomization tool generated sets of two numbers (1 and 2, translated to conditions A and B), randomly ordered. With the Figure 1C data, in which $n_A = n_B = 13$, this leads to 8192 randomizations, which represent 8192 sequences of the 13 A and 13 B labels that could have been obtained with the random assignment procedure followed, which means that the test statistic of choice is to be applied 8192 times to the same data sequence, which under the null hypothesis could have been obtained regardless of the conditions in each measurement occasion. (The same interpretation for the number randomizations is warranted for the remaining applications of randomization tests presented in this section.) The randomization test applied with the SCDA plug-in for R (Bulté & Onghena, 2013) yields a one-tailed p = .000244 for the simple mean difference and the R code we developed for ALIV yields a one-tailed p = .000122 for ALIV.

In case the data show overlap and different slopes and intercepts for the different conditions (e.g., the Yakubova and Bouck, 2014, data for Rick represented on Figure 1C; illustration 3 from the online supplementary material), the mean difference yields the greatest value and linear regression yields the smallest one, but these are also the procedures that represent the data worse. The remaining procedures, which show less extreme results, actually provide a better fit to the data. The randomization scheme followed was based on flipping a coin to decide which condition takes place when, with a restriction of a maximum of two consecutive sessions with the same condition. The actual data consist of the same number of measurements per condition ($n_A = n_B = 5$), but it is not clear whether this was decided a priori. Therefore, we will illustrate the application of a randomization test for the following example in which such a specification is available.

In case the data show different intercepts and opposite slopes for the different conditions (e.g., the Sil, Dahlquist, and Burns, 2013, data for child cooperation as reported by the nurses represented on Figure 1D; illustration 4 from the online supplementary material), all procedures yield similar average differences, as both linear and quadratic trend fit the data reasonably well. In this case, the mean levels do not represent the data well and the mean difference provides the smallest value. The projection made by piecewise regression for the last measurement occasions does not seem to be justified, given the large (observed and expected) difference. For this data set it is relevant to note that ADISO assigns more weight to the first value (47.6) of the interactive distraction condition (filled triangle) and to the last value (19.7) of the passive distraction condition (empty square), given that these are the only data points for the corresponding condition, surrounded by four measurements of the other condition. This weighting scheme is reasonable, because the isolated data points are crucial for the comparison

between conditions. Complementarily, ALIV and ADISO assign less weight to data points which cannot be compared with a contiguous measurement from the other condition, in particular, the first measurement of the passive distraction condition (empty square) and the last measurement of the interactive distraction condition (filled triangle). We consider that these weights correspond more closely to the assessment likely to be performed by visual analysts, who would compare the lines connecting the points belonging to the same condition. Therefore, the focus in visual analysis and in ALIV is likely to be placed on the same portion of the data. In terms of the application of a randomization test, in the Sil et al. (2013) study, a semi-random order is used, with 10 measurement occasions, 5 per each condition, and no more than 2 consecutive applications of the same condition. This leads to 84 possible randomizations. The randomization test applied with the SCDA plug-in for R (Bulté & Onghena, 2013) yields a one-tailed $p = .0952$ for the simple mean difference and the R code we developed for ALIV yields a one-tailed $p = .1071$ for ALIV.

In case the data show very different trends in different conditions, including linear, quadratic, and another difficult to identify trend (e.g., the Bryant et al., 2015, data for John represented on Figure 1E; illustration 5 from the online supplementary material), the results of the procedures quantifying average difference agree less than for the previous data patterns. The mean level, piecewise, linear and quadratic regression models represent either one condition or all conditions insufficiently well. When the data for the different conditions are so diverse, neither of these methods is recommended. LOESS and ALIV provide very similar results for the fractions of data to which the local regressions are applied, although this is not necessarily certain for other fraction parameters for LOESS. The results of ADISO are very different, probably in relation to the specific segmentation chosen. In case the segmentation of the data sequence is not clear, the

variation of results according to the segmentation is a drawback, given that trying several different options is a time-consuming task. Note that some of the comparisons for PND-W entail very distant measurement occasions (e.g., the second and the seventh, which represent the second data point for conditions 1 and 3, respectively), which may not be justified. The randomization scheme followed in the Bryant et al. (2015) study consisted in randomly determining the sequence of three treatments, each appearing five times over a period of 15 measurement occasions. From the text it appears that no further restrictions were imposed (i.e., a completely randomized design is followed), evidence for which is the fact that one of the conditions represented on Figure 1E is present on three consecutive measurement occasions. Thus, there are $15!/(5!\,5!\,5!) = 756\,756$ possible random orders. For pairwise comparisons between conditions, there would be $10!/(5!\,5!) = 252$ possible random orders. However, for John, whose data is depicted on Figure 1E, there are only 13 measurement occasions, with one condition appearing 5 times and the remaining two 4 times each, leading to $13!/(5!\,4!\,4!) = 90\,090$ random orders for the three conditions and $8!/(4!\,4!) = 70$ or $9!/(45!\,4!) = 126$ random orders for the pairwise comparisons. Due to space limitations we do not present these pairwise comparisons here.

In case the data show a large degree of overlap and nonlinear trends (e.g., the Eilers and Hayes, 2015, data for Jacob represented on Figure 1F; illustration 6 from the online supplementary material), mean level, piecewise, linear and quadratic regression models provide poor fit to the data. The mean difference is the procedure that shows most distant results. Even the LOESS model chosen does not provide fit as good as for the previous examples, suggesting that the same fraction and the same degree of polynomial may not be useful for all data sets. The segmentation chosen for ADISO provides slightly larger values than ALIV and the regression-

based models, but the percentage of comparisons for which one condition is superior to the other is consistent with the value of PND-W. However, the PND-W omits the last two data points, as $n_A \neq n_B$, and the first two comparisons are actually not between adjacent values. Comparing the new proposals with the simple mean difference, ALIV and ADISO assign less weight to the first two and last two measurements. We consider that this weighting scheme is more appropriate, given that these initial and final values represent repetitions of the same condition without the possibility of knowing what the results would have been in case condition B (filled triangle) took place before A (empty circle) in the beginning of the sequence or in case condition A took place after condition B in the end of the sequence, because it is not reasonable to extend the clearly nonlinear trends beyond the measurements occasions for which data were actually obtained. Moreover, the graphical representation provided jointly with ALIV illustrates better than the simple mean difference the fact that the difference between conditions is not uniformly in the same direction or in the same magnitude, if interpolated values are considered. In that sense, if it is judged that there is a difference between conditions which is increasing with time (i.e., more visible for the later part of the data sequence), such information is more clearly illustrated by the ALIV graph and quantifications than by the simple mean difference. Even if the focus is put only on the actually obtained values, as in ADISO, the differences between conditions are still illustrated to be clearly variable and not even systematically increasing or decreasing. Thus, the information provided by ALIV and ADISO is more specific, as compared to the simple mean difference.

In terms of the application of a randomization test in the Eilers and Hayes (2013) study, a semi-random order was used, with 16 measurement occasions, and no more than 2 consecutive applications of the same condition. However, in this case there is apparently no restriction about

both conditions being equally represented, as in the actual data sequence there are seven measurement occasions condition A and nine for condition B. If we focus on designs in which both measurements are equally represented, there would be 1296 possible random assignments (obtainable from the SCDA software; Bulté & Onghena, 2013) plus 786 possible randomizations in case condition A had seven measurements and condition B had nine plus 786 possible randomizations in case condition B had seven measurements and condition A had nine (obtainable via the executable files available in the CD accompanying the book by Edgington & Onghena, 2007). The randomization test performed via the SCDA plug-in for R (Bulté & Onghena, 2013) on the basis of these 2868 randomizations yields one-tailed $p = .0146$ for the mean difference and the R code created for ALIV yields one-tailed $p = .0948$, which would lead to different statistical decisions being made on the basis if the common .05 alpha level.

In sum, the examples shown in the present section suggest that ALIV and ADISO can be applied to a variety of data sets, presenting the following positive features: (a) a good representation of the data (unlike the mean level and linear or quadratic regression which oversimplify certain data patterns), (b) no need for specifying a priori the type of trend (unlike piecewise regression) or assuming it is absent (unlike the mean difference), (c) comparisons between values that are close in the sequence (unlike PND-W in some cases), (d) no need for making decisions about how well the model should fit the data and no need for varying modelling parameters for each specific data set (unlike LOESS), (e) provide an overall quantification without the need for collating the information about slope and the comparison for a single measurement occasion (unlike piecewise regression), (f) ADISO-O can also provide information similar to PND-W, but without the restriction for having the same number of measurement occasions in all conditions, and (g) a graphical representation that provides

information about the differences between the conditions at different points of the sequence, apart from yielding an overall quantification.

It may be argued that, given the similarity in results (in some cases), the mean difference could be preferred as a simpler and more parsimonious option to either ALIV or ADISO, when justified. However, given that the new proposals (a) can be applied with the free R code, (b) entail more meaningful and more specific comparisons mimicking visual analysis (ALIV) or focusing on contiguous conditions (ADISO), and (c) represent the data better for a wide variety of data patterns, we consider that they should be the preferred option. Nevertheless, the potential limitations of ALIV and ADISO mentioned in Table 2 need also be taken into account. Moreover, simulation studies would be useful to show whether ALIV or ADISO present a statistical power advantage over the mean difference when used as a test statistic in a randomization test.

Discussion

How can ATD data be analyzed? Analytical Techniques Reviewed and New Proposals

The possibility to identify evidence-based practice through SCEDs is related to both using appropriate design structures (Kratochwill et al., 2013) and summarizing the results of the studies with adequate quantifications of intervention effect (Jenson, Clark, Kircher, & Kristjansson, 2007). In the context of ATDs, determining the alternations at random provides another basis for obtaining solid evidence, apart from ensuring a sufficient number of comparisons between conditions. In terms of analyzing the data and summarizing the results to make them available for documenting treatment effects, in the first part of the paper we discussed existing techniques suggested for application to ATD designs in relation to the specific

features of these designs that distinguish them from phase designs. We also reviewed recent published research to identify which of these techniques have been most commonly used.

On the one hand, the most commonly used analytical strategy in the published research reviewed was visual analysis, which enables taking into consideration several features of the data, but the application of the analytical steps detailed in Kratochwill et al. (2010) is not as straightforward for ATDs and, additionally, comparison and integration of results across studies is limited. On the other hand, the most common quantification used is mean difference between conditions, accompanied by reporting a measure of data variability in each condition. Both these quantifications, as well as the third most used one (PND), do not take trend into account. In that sense, we want to raise awareness about the importance of trend, which has been found to be present in real data, despite being heterogeneous across studies (Solomon, 2014), and also encountered, in a linear or nonlinear form, in most studies included our review. Actually, controlling for trend has received a lot of attention when discussing SCED analytical techniques (Parker et al., 2006) and is part of simple procedures such as graph rotation (Parker, Vannest, & Davis, 2014), nonoverlap indices (Wolery, Busick, Reichow, & Barton, 2010) and more complex techniques such as multilevel models (Moeyaert, Ferron et al., 2014). In that sense, the current paper fills a gap in SCED analysis literature regarding ATDs, especially given that our review suggests that little attention is paid to trend when analyzing ATD data.

In relation to enabling comparisons between conditions without assuming any specific data pattern, local regression has been suggested, but it entails subjective decisions leading to the model finally selected and its application may also be problematic for the relatively short data series encountered in our review of published research. In order to deal with these limitations, and still not assuming any specific data pattern, we propose a technique comparing actual and

linearly interpolated values (ALIV). ALIV quantifies, for each measurement occasion, the differences between the lines connecting the values for the same condition, which are the common way of representing ATD data graphically.

Apart from possible linear or nonlinear trends, another relevant data feature in ATDs is that the number of measurements per condition is not always equal, as suggested by our review. This limits the applicability of PND-W, which also does not guarantee that adjacent values are being compared. In order to deal with these limitations, we propose a technique for computing the average difference between successive observations (ADISO). ADISO also offers the possibility to quantify the amount of difference between the conditions, apart from superiority as quantified by PND-W.

How should ATD data be analyzed? Considerations regarding Statistical Analysis

Both for ALIV (as alternative to the simple mean difference and to local regression) and for ADISO (as alternative to the simple mean difference and to PND-W), we have stressed the importance of what is conceptually being compared in an ATD rather than the numerical differences between alternative analyses, as well as the need to focus on more generally applicable procedures that make less assumptions and require fewer subjective decisions by the researchers. Specifically, we consider that it is more logical to compare adjacent conditions (Graham, Karmarkar, & Ottenbacher, 2012), as is also suggested for other SCED designs (Maggin et al., 2013; Parker & Vannest, 2012). In that sense, ADISO quantifies the effect in a way that is more compatible with the ATD structural logic than the overall within-condition means. Additionally, ADISO-O quantifying the percentage of comparisons for which one of the

conditions is superior is an indicator of the degree to which the effect is consistent across all repetitions, as suggested by Kratochwill et al. (2010).

We also consider that ATDs offer the possibility to construct the counterfactual of a measurement (i.e., what value would have been obtained if the other condition were administered on that occasion) via the randomization model and/or the appropriate regression model. An additional option is to carry out interpolations that refer to a measurement occasion in between the occasions for which data are available, which means that unreasonable projections are not that likely as for phase designs where projections may refer to some distant point in time (see Parker, Vannest, Davis, & Sauber's, 2011, comments on Allison & Gorman's, 1993, regression model projections). In that sense, ALIV reflects the spirit of interpolating and allows modeling data on the basis of the neighboring values, bringing the statistical analysis closer to the visual inspection of the data. Specifically, the size of arrows as visual aids indicates difference in level (see illustration 6 in the online supplementary material), whereas in case the arrows become longer in time it would indicate difference in slope (illustration 2). Moreover, the direction and color of the arrows help assessing overlap (illustration 6) or lack thereof (illustration 2). Finally, ALIV entails a comparison between actual and projected data, as suggested in the What Works Clearinghouse *Standards* (Kratochwill et al., 2010).

The fact that the alternation sequence is frequently determined at random in an ATD makes possible the application of randomization tests. The test statistic can be chosen according to the predicted effect and the expected data pattern. We consider ALIV to be more generally applicable across data sets than the mean difference, which assumes data without trend.

An aspect that needs to be taken into account when choosing an analytical technique is whether it is preferred to use quantifications that are generally applicable across a variety of SCEDs (e.g., the mean difference) or quantifications that have been created specifically for certain designs, as is the case of ADISO and ALIV. The former option offers an advantage at the across-studies level, as it ensures comparability of results and the possibility to integrate the results of SCED studies regardless of the specific design used. Nevertheless, we consider that options such as ADISO and ALIV are more informative at the within-study level as they quantify aspects of interest (i.e., adjacent values in ADISO, the lines connecting the values belonging to different conditions in ALIV) and they are also meaningful regardless of whether any type of trend is present in the data. Moreover, the fact that the same quantification (e.g., a standardized mean difference) can be obtained across designs does not ensure that it is conceptually justified to pool the information, as the behaviors and interventions subjected to ATDs can be expected to be different from the ones studied via phase designs and especially multiple-baseline designs, employed, among other reasons, for nonreversible behaviors (in contrast with ATDs).

Finally, if the new proposals that we advocate for are to be used by applied researchers, hand calculations are possible, but may become cumbersome for longer data series. For that purpose we offer a user-friendly webpage (http://manolov.shinyapps.io/ATDesign) and R code (indications on the use of the latter are available in the supplementary online material). Its use is based on, first, inputting the data and, afterwards, copying and pasting the code in the R console. Part of the output are data plots in which the color of the graphical elements indicates whether the difference observed in each of the ADISO and ALIV comparisons is in the same or in the

opposite direction to what is desired. R code is also offered for the existing analytical techniques discussed in this manuscript.

Considerations regarding the Joint Use of Visual and Statistical Analyses

On the basis of the alternatives presented in the previous section, we concur with Parker et al. (2006) that visual analysis is a necessary part of the analysis in order to carry out a general assessment of the data pattern and also to evaluate specifically whether the data series are stationary or present trends, and to identify any outlying values. It has already been stressed that visual and statistical analyses need to be seen as complementary (Fisch, 2001; Franklin, Gorman, Beasley, & Allison, 1996; Harrington & Velicer, 2015; Houle, 2009), for instance, in relation to the importance of making SCED studies' results available for meta-analysis (Burns, 2012; Maggin & Chafouleas, 2013). However, it has to be considered whether visual analysis has to be carried out as an *initial* step that determines subsequent descriptive or inferential statistical analyses or whether it should be one of the ingredients that helps making sense of the data obtained.

Some researchers who argue to carry out visual analysis as an initial step use this kind of analysis as a filter indicating whether statistical analysis is necessary. If statistical analysis is performed only after visual analysis suggests a functional relation between conditions and the target behavior (as recommended by Kratochwill et al. 2010, 2013), then it can be expected that the numerical values obtained would only represent part of the empirical evidence and a subsequent meta-analysis would include a biased sample of all research carried out (i.e., larger effects). If statistical analysis is performed only after visual analysis is unable to identify a clear-cut effect or when visual analysis can be considered unreliable (as recommended by Kazdin,

1978), then the numerical values would only represent part of the empirical evidence – in this case the research with smaller effects – leading to a biased sample in posterior meta-analysis. An alternative would be to always use statistical analysis and visual analysis jointly.

The initial step could also refer to the idea that visual analysis can inform us about the most salient features of the data and thus help choosing the statistical technique that would represent such data best. However, such a two-stage procedure does not guarantee that the Type I error is controlled at the nominal level. It might nearly always be possible to find some distinction between the measurements from different conditions, potentially capitalizing on chance. After data transformation or data fitting it might nearly always be possible to find intervention effects, defined as either change in level, or in slope, or in variability, or in the amount of overlap (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). An alternative way of proceeding would be to select a way of analyzing the data according to the type of effect expected before the data are collected or before the results are revealed (e.g., according to whether an abrupt and sustained or a slower and more progressive change is expected for a specific behavior treated with a specific intervention). Afterwards, visual analysis can be used to give meaning to the obtained results and to assess to what extent the data at hand match previous expectations. Such evidence can be used for adjusting the expectations for future research.

This issue is similar to the concern expressed, albeit in a footnote, by Kratochwill and Levin (2014a) that repeatedly adapting statistical models to represent the data and to estimate the intervention effect can lead to specifying models without further basis than the data at hand, which is a problem both in terms of the causal inferences that can be made and in terms of the ethics of data analysis. We extend Kratochwill and Levin's (2014a) concern beyond the application of multilevel modelling to SCED data, as we consider that a similar caution has to be

expressed when choosing the type of analysis to be carried out on the basis of the data at hand and not on theoretical and/or empirical grounds (see also Simmons, Nelson, & Simonsohn, 2011). Additionally, a parallelism can be drawn between, on the one hand, the increase in Type I errors in response-guided experiments in which the incoming data are assessed continuously and used to make decisions about the changes in the conditions (Ferron & Jones, 2006) and, on the other hand, the problems that can arise from a conditional or repetitive process of data analysis in response to the data at hand. Similarly, just like it has been suggested not to ground the analytical process on the specific characteristics of the data, it has also been suggested not to decide when to end data collection on the basis of the data themselves (Howard, Best, & Nickels, 2015).

Moreover, the concerns expressed here are also well-aligned with the broader statistical literature on the effects of using preliminary tests (in the current case, performed visually) for deciding for choosing the predictors in regression analysis (increase in bias of the regression coefficients estimates; Bancroft, 1944) or whether to use a pooled estimate of the variance or not (increase in bias of the variance estimate; Bancroft, 1944, increase in Type I error rates, Zimmerman, 2004). It has also been pointed out that, in some realistic situations, the imperfect performance of the first-stage test can lead to worse results for the main second-stage test (Shuster, 2005). This could be extended to the imperfect performance of visual analysts (e.g., Danov & Symons, 2008; Ninci, Vannest, Willson, & Zhang, 2015) as a possible first-step in the data analysis process. Another recommendation from the general statistical literature is to use analytical procedures whose validity is not based on assuming or testing for specific data features: randomizations tests were specifically mentioned (Schucany & Ng, 2006) and we add ALIV and the weighted ADISO as another alternative following this recommendation.

To summarize, we consider that SCED data analysis is not that different from data analysis in general. Despite the specific characteristics of SCED data (i.e., repeated and potentially serially dependent measurements from a single unit under different conditions), the same families of analyses have been suggested: standardized mean difference (Busk & Serlin, 1992; Hedges et al., 2012, 2013), regression-based models (Allison & Gorman, 1993; Swaminathan et al., 2014) and hierarchical linear models as extensions of the piecewise regression (Moeyaert, Ferron et al., 2014), randomization tests (Edgington & Onghena, 2007) and even some of the apparently SCED-specific nonoverlap indices are closely related or equivalent to effect size measures used in between-groups designs (see NAP; Parker & Vannest, 2009 and Tau-U; Parker, Vannest, Davis, & Sauber, 2011). Therefore, the considerations regarding two-stage data analysis made outside of the SCED context can also be extendable to SCEDs, especially in relation to how visual and statistical analysis are applied together on the same data.

Considerations regarding Causal Relations

In the present article we stressed the importance of assessing the effects in ATDs not only in terms of effect sizes for the outcome measure, but also according to the characteristics of the design, such as the way in which the conditions are alternated (preferably at random), the number of repetitions of the alternations, the number of measurements per condition, and the spacing between sessions. The appropriateness of the design for demonstrating a causal relation between the type of condition and the behavior of interest is an initial requirement (Kratochwill et al., 2010), whereas the visual inspection of the data can help assessing whether such a relation has actually been demonstrated. Focusing on four of the six data aspects highlighted in the What Works Clearinghouse *Standards* (the two others: within-condition variability and immediacy of the effect are less straightforward criteria in presence of rapid alternation of conditions), visual

analysis, visual aids, and quantifications such as the means, regression intercepts and slopes, and nonoverlap indices can all be useful. Visual analysis is especially useful for evaluating the general data pattern and assessing the consistency between the measurements obtained in different conditions and the changes between conditions, whereas the comparison between fitted values arising from regression analysis and ALIV is especially useful as an approximation to the comparison between actual and projected data. Actually, regression analysis can be useful for exploring the type of functional relation between time and measurements within each condition (e.g., a stationary process, linear, quadratic or more complex model).

Another consideration when assessing causal effects is that both the effect size for the outcome variable and the size of the manipulation in the independent variable have to be taken into account. Or in other words: the effect size has to be evaluated proportional to the manipulation size. Looking at the design and the effects in this way, even small effects may be impressive (Prentice & Miller, 1992). Furthermore, if we want to know the "functional relation" between the independent variable and the outcome variable, as is implied in the What Works Clearinghouse *Standards* and some of the commentaries (Hitchcock et al. 2014; Kratochwill et al., 2010, 2013; Maggin, 2015; Maggin, Biesch, & Chafouleas, 2013; Wolery, 2013), then we have to do more than just study the effects of a dichotomous independent variable or exploring the functional relation between time and the outcome variable. If we want to map the functional relation between the independent variable and the outcome variable, then even more firm evidence for a causal effect can be obtained by systematically varying the levels of the independent variable, for example in a so-called parametric variation design (Barlow et al., 2009; Kazdin, 2011; Kratochwill & Levin, 2014b; Onghena, 1992).

Limitations and Future Research

Regarding the limitations of the current article, we here presented, discussed and illustrated some analytical alternatives for ATDs, whereas a formal evaluation via simulation should be done in subsequent research. Specifically, the evaluation can focus on how well the different regression models represent ATD data with and without several types of trend for varying amount of measurements available per condition. To inform about the range of measurements usually available in an ATD and that need to be represented in a simulation study, the information from previous reviews (e.g., Shadish & Sullivan, 2011) can be used, as well as the more specific information obtained in the review included in the present paper.

In terms of data analysis for ATDs, further software implementations for ADISO are necessary in order to make its use as a test statistic in a randomization test feasible. Moreover, additional discussion and research is required regarding the analysis of ATDs beyond the comparison phase of rapid alternation of conditions, that is, when there is an initial baseline phase and/or a final phase where only the best intervention is implemented.

References

*The articles marked with an asterisk are the ones included in the review of published alternating treatment designs research.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621–631.

*Alstot, A. E. (2012). The effects of peer-administered token reinforcement on jump rope behaviors of elementary physical education students. *Journal of Teaching in Physical Education, 31,* 261–278.

*Andersen, M. N., Daly, III, A. J., & Young, N. D. (2013). Examination of a one-trial brief experimental analysis to identify reading fluency interventions. *Psychology in the Schools, 50*, 403–414.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics, 15*, 190–204.

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12,* 199–210.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

*Begeny, J. C., Yeager, A., & Martínez, R. S. (2012). Effects of small-group and one-on-one reading fluency interventions with second grade, low-performing Spanish readers. *Journal of Behavioral Education, 21*, 58–79.

*Bellone, K. M., Dufrene, B. A., Tingstrom, D. H., Olmi, D. J., & Barry, C. (2014). Relative efficacy of behavioral interventions in preschool children attending head start. *Journal of Behavioral Education, 23,* 378–400.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129-141.

*Bickford, J. O., & Falco, R. A. (2012). Technology for early braille literacy: Comparison of traditional braille instruction and instruction with an electronic notetaker. *Journal of Visual Impairment & Blindness, 106*, 679–693.

Borckardt, J., & Nash, M. (2014). Simulation modelling analysis for small sets of single-subject data collected over time. *Neuropsychological Rehabilitation, 24*, 492–506.

Brossart, D. F., Vannest, K., Davis, J., & Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation, 24*, 464–491.

*Bryant, B. R., Ok, M., Kang, E. Y., Kim, M. K., Lang, R., Bryant, D. P., & Pfannestiel, K. (2015). Performance of fourth-grade students with learning disabilities on multiplication facts comparing teacher-mediated and technology-mediated interventions: A preliminary investigation. *Journal of Behavioral Education, 24*, 255–272.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 104–114.

Bulté, I., & Onghena, P. (2013). The Single-Case Data Analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods, 12*, 450–478.

Burns, M. K., (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education, 21*, 175-184.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum.

*Capriotti, M. R., Brandt, B. C., Ricketts, E. J., Espil, F. M., & Woods, D. W. (2012). Comparing the effects of differential reinforcement of other behavior and response-cost contingencies on tics in youth with Tourette syndrome. *Journal of Applied Behavior Analysis, 45*, 251–263.

*Carroll, R. A., Kodak, T., & Adolf, K. J. (2016). Effect of delayed reinforcement on skill acquisition during discrete-trial instruction: implications for treatment-integrity errors in academic settings. *Journal of Applied Behavior Analysis*, *49*, 176-181.

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data what they tell us and what they don't. *Behavior Modification, 37*, 378-390.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*, 387–400.

*Chan, J. M., O'Reilly, M. F., Lang, R. B., Boutot, E. A., White, P. J., Pierce, N., & Baker, S. (2011). Evaluation of a Social Stories™ intervention implemented by pre-service teachers for

students with autism in general education settings. *Research in Autism Spectrum Disorders, 5*, 715–721.

Cohen, R. A. (n.d). *An Introduction to PROC LOESS for Local Regression*. Retrieved November 2, 29015 from http://statcomp.ats.ucla.edu/stat/sas/library/loesssugi.pdf

*Coleman, M. B., Cherry, R. A., Moore, T. C., Park, Y., & Cihak, D. F. (2015). Teaching sight words to elementary students with intellectual disability and autism: A comparison of teacher-directed versus computer-assisted simultaneous prompting. *Intellectual and Developmental Disabilities, 53*, 196–210.

*Couper, L., van der Meer, L., Schäfer, M. C., McKenzie, E., McLay, L., O'Reilly, M. F., ... & Sutherland, D. (2014). Comparing acquisition of and preference for manual signs, picture exchange, and speech-generating devices in nine children with autism spectrum disorder. *Developmental Neurorehabilitation, 17*, 99–109.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*, 828–839.

*Devlin, S., Healy, O., Leader, G., & Hughes, B. M. (2011). Comparison of behavioral intervention and sensory-integration therapy in the treatment of challenging behavior. *Journal of Autism and Developmental Disorders, 41*, 1303–1320.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small*-n *experimental designs: A practical guide to randomization tests* (2nd ed.). New York, NY: Routledge.

Edgington, E. S. (1967). Statistical inference from N=1 experiments. *Journal of Psychology, 65*, 195–199.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80*, 351−363.

Edgington, E. S. (1980a). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2,* 19–28.

Edgington, E. S. (1980b). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235–251.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34,* 567–574.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

*Eilers, H. J., & Hayes, S. C. (2015). Exposure and response prevention therapy with cognitive defusion exercises to reduce repetitive and restrictive behaviors displayed by children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 19*, 18–31.

Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review, 14,* 532–550.

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation, 24*, 305–314.

Evers, C. W., & Wu, E. H. (2006). On generalising from single case studies: Epistemological reflections. *Journal of Philosophy of Education, 40,* 511–526.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Greenwich, CT: Information Age Publishing.

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66–81.

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153–183). Washington, DC: American Psychological Association.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54*, 137–154.

*Fletcher, D., Boon, R. T., & Cihak, D. F. (2010). Effects of the TOUCHMATH program compared to a number line strategy to teach addition facts to middle school students with moderate intellectual disabilities. *Education and Training in Autism and Developmental Disabilities*, *45*, 449–458.

*Flosason, T. O., McGee, H. M., & Diener-Ludwig, L. (2015). Evaluating impact of small-group discussion on learning utilizing a classroom response system. *Journal of Behavioral Education, 24*, 1–21.

*Flower, A. (2014). The effect of iPad use during independent practice for students with challenging behavior. *Journal of Behavioral Education, 23*, 435–448.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12,* 219–245.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison & B. S. Gorman (Eds*.), Design and analysis of single-case research* (pp. 119–158). Mahwah, NJ: Lawrence Erlbaum.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology, 25*, 46–60.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199–233). London, UK: Routledge.

Graham, J. E., Karmarkar, A. M., & Ottenbacher, K. J. (2012). Small sample research designs for evidence-based rehabilitation: Issues and methods. *Archives of Physical Medicine and Rehabilitation, 93*, S111–S116.

Guyatt, G. H., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The n-of-1 randomized controlled trial: Clinical usefulness - our three-year experience. *Annals of Internal Medicine, 112,* 293–299.

Guyatt, G., Jaeschke, R., & McGinn, T. (2002). PART 2B1: Therapy and validity. N-of-1 randomized controlled trials. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.), *Users' guides to the medical literature* (pp. 275–290). New York, NY: McGraw-Hill.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities, 45*, 187–202.

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*, 162–183.

*Hastie, P., van der Mars, H., Layne, T., & Wadsworth, D. (2012). The effects of prompts and a group-oriented contingency on out-of-school physical activity in elementary school-aged students. *Journal of Teaching in Physical Education, 31*, 131–145.

*Hawkins, R. O., Hale, A., Sheeley, W., & Ling, S. (2011). Repeated reading and vocabulary-previewing interventions to improve fluency and comprehension for struggling high-school readers. *Psychology in the Schools, 48*, 59–77.

*Haydon, T., Conroy, M. A., Scott, T. M., Sindelar, P. T., Barber, B. R., & Orlando, A. M. (2010). A comparison of three types of opportunities to respond on student academic and social behaviors. *Journal of Emotional and Behavioral Disorders, 18*, 27–40.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods, 1*, 184-198.

*Heinicke, M. R., Carr, J. E., Eastridge, D., Kupfer, J., & Mozzoni, M. P. (2013). Assessing preferences of individuals with acquired brain injury using alternative stimulus modalities. *Brain Injury, 27*, 48–59.

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science, 3,* 51–64.

Heyvaert, M., Wendt, O., Van Den Noortgate, W., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *Journal of Special Education, 49,* 146–156.

Hinkelmann, K., & Kempthorne, O. (2008). *Design and analysis of experiments, Volume 1: Introduction to experimental design* (2nd ed.). New York, NY: Wiley.

Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2014). The What Works Clearinghouse single-case design pilot standards: Who will guard the guards? *Remedial and Special Education, 35,* 145–152.

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single subject research: Description of designs and discussion of problems. *Topics in Early Childhood and Special Education, 16*, 168–190.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (3rd Ed.) (pp. 271–305). Boston, MA: Pearson.

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: critiques and ways forward. *Aphasiology, 29*, 526–562.

Hurvich, C. M., Simonoff, J. S., & Tsai, C-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 60*, 271–293.

Jacoby, W. G. (2000). Loess: A nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies, 19*, 577–613.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483–493.

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*, 629–642.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152.

Kempthorne, O. (1979). Sampling inference, experimental inference and observation inference. *Sankhyā: The Indian Journal of Statistics, Series B, 40,* 115–145.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

*Klubnik, C., & Ardoin, S. P. (2010). Examining immediate and maintenance effects of a reading intervention package on generalization materials: Individual verses group implementation. *Journal of Behavioral Education, 19*, 7–29.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26–38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 124–144.

Kratochwill, T. R., & Levin, J. R. (2014a). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology, 52*, 231–235.

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014b). *Single-case intervention research: Statistical and methodological advances.* Washington, DC: American Psychological Association.

*Lang, R., Davis, T., O'Reilly, M., Machalicek, W., Rispoli, M., Sigafoos, J., ... & Regester, A. (2010). Functional analysis and treatment of elopement across two school settings. *Journal of Applied Behavior Analysis, 43*, 113–118.

*Lang, R., O'Reilly, M., Sigafoos, J., Machalicek, W., Rispoli, M., Lancioni, G. E., ... & Fragale, C. (2010). The effects of an abolishing operation intervention component on play skills, challenging behavior, and stereotypy. *Behavior Modification, 34*, 267–289.

*Lang, R., Rispoli, M., Sigafoos, J., Lancioni, G., Andrews, A., & Ortega, L. (2011). Effects of language of instruction on response accuracy and challenging behavior in a child with autism. *Journal of Behavioral Education, 20*, 252–259.

*Lee, G. T., & Singer-Dudek, J. (2012). Effects of fluency versus accuracy training on endurance and retention of assembly tasks by four adolescents with developmental disabilities. *Journal of Behavioral Education, 21*, 1–17.

Levin, J. R., Evmenova, A. S., & Gafurov, B. S. (2014). The single-case data-analysis ExPRT (Excel Package of Randomization Tests). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 185–219). Washington, DC: American Psychological Association.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB…AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50*, 599-624.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Losinski, M., Maag, J. W., Katsiyannis, A., & Ryan, J. B. (2015). The Use of Structural Behavioral Assessment to Develop Interventions for Secondary Students Exhibiting Challenging Behaviors. *Education and Treatment of Children, 38*, 149–174.

Maggin, D. M. (2015). Considering generality in the systematic review and meta-analysis of single-case research: A response to Hitchcock et al. *Journal of Behavioral Education*, *24*, 470–482.

Maggin, D. M., Biesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*, 44–58.

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the Special Series: Issues and advance of synthesizing single-case research. *Remedial and Special Education, 34*, 3-8.

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior Modification, 38*, 878–913.

*McLay, L., van der Meer, L., Schäfer, M. C., Couper, L., McKenzie, E., O'Reilly, M. F., ... & Sutherland, D. (2015). Comparing acquisition, generalization, maintenance, and preference across three AAC options in four children with autism spectrum disorder. *Journal of Developmental and Physical Disabilities, 27*, 323–339.

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2016, February 29). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods*. Advance online publication. doi: 10.3758/s13428-016-0714-4

Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology, 52*, 191–211.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van Den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behavior Modification, 38*, 665–704.

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., Beretvas, S., & Van Den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly, 30,* 50–63.

*Mong, M. D., & Mong, K. W. (2010). Efficacy of two mathematics interventions for enhancing fluency with elementary students. *Journal of Behavioral Education, 19*, 273–288.

*Mong, M. D., & Mong, K. W. (2012). The utility of brief experimental analysis and extended intervention analysis in selecting effective mathematics interventions. *Journal of Behavioral Education, 21*, 99–118.

*Morgan, R. L., & Horrocks, E. L. (2011). Correspondence between video-based preference assessment and subsequent community job performance. *Education and Training in Autism and Developmental Disabilities, 46*, 52–61.

Morgan, D. L., & Morgan, R. K. (2009). *Single-case research methods for the behavioral and health sciences*. Thousand Oaks, CA: Sage.

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification, 39*, 510–541.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology, 7*, e934.

Olive, M. L., & Franco, J. H. (2008). (Effect) size matters: And so does the calculation. *The Behavior Analyst Today, 9*, 5–10.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*, 783–786.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*, 56–68.

Pane, H. M., Sidener, T. M., Vladescu, J. C., & Nirqudkar, A. (2015). Evaluating function-based Social Stories™ with children with autism. *Behavior Modification, 39*, 912–931.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418–443.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357–367.

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*, 254–265.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303–322.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *Journal of Special Education, 48*, 79–91.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*, 284–299.

*Petursdottir, A. I., & Aguilar, G. (2016). Order of stimulus presentation influences children's acquisition in receptive identification tasks. *Journal of Applied Behavior Analysis, 49*, 58-68.

Pfadt, A., Cohen, I., Sudhalter, V., Romanczyk, R., & Wheeler, D. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis, 25*, 551–560.

*Plavnick, J. B., & Ferreri, S. J. (2011). Establishing verbal repertoires in children with autism using function-based video modeling. *Journal of Applied Behavior Analysis, 44*, 747-766.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

*Reed, D. D., Luiselli, J. K., Morizio, L. C., & Child, S. N. (2010). Sequential modification and the identification of instructional components occasioning self-injurious behavior. *Child & Family Behavior Therapy, 32*, 1–16.

*Rich, S. E. H., & Duhon, G. J. (2014). Using brief academic assessments to determine generalization strategies. *Journal of Behavioral Education, 23*, 401–420.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185–193.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100,* 322–331.

*Sabielny, L. M., & Cannella-Malone, H. I. (2014). Comparison of prompting strategies on the acquisition of daily living skills. *Education and Training in Autism and Developmental Disabilities, 49*, 145–152.

*Schneider, A. B., Codding, R. S., & Tryon, G. S. (2013). Comparing and combining accommodation and remediation interventions to improve the written-language performance of children with Asperger syndrome. *Focus on Autism and Other Developmental Disabilities, 28*, 101–114.

Schucany, W. R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student *t*. *Communications in Statistics - Theory and Methods, 35*, 2275–2286.

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*, 9-19.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109–122.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*, 123–147.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385–405.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188–196.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971–980.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology, 52*, 149–178.

Shuster, J. (2005). Diagnostics for assumptions in moderate to large simple trials: Do they really help? *Statistics in Medicine, 24*, 2431–2438.

*Sil, S., Dahlquist, L. M., & Burns, A. J. (2013). Case study: videogame distraction reduces behavioral distress in a preschool-aged child undergoing repeated burn dressing changes: a single-subject design. *Journal of Pediatric Psychology, 38*, 330–341.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366.

*Simonsen, B., MacSuga, A. S., Fallon, L. M., & Sugai, G. (2013). The effects of self-monitoring on teachers' use of specific praise. *Journal of Positive Behavior Interventions, 15*, 5–15.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550.

Solmi, F., Onghena, P., Salmaso, L., & Bulté, I. (2014a). A permutation solution to test for treatment effects in alternation design single-case experiments. *Communications in Statistics - Simulation and Computation, 43*, 1094–1111.

Solmi, F., Onghena, P., Salmaso, L., & Bulté, I. (2014b). Extensions of permutation solutions to test for treatment effects in replicated single-case alternation experiments with multivariate response. *Communications in Statistics: Simulation and Computation, 43*, 1036–1051.

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*, 477–496.

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education, 24*, 438-458.

*Sprinkle, E. C., & Miguel, C. F. (2013). Establishing derived textual activity schedules in children with autism. *Behavioral Interventions, 28*, 185–202.

*Steiner, A. M. (2011). A strength-based approach to parent education for children with autism. *Journal of Positive Behavior Interventions, 3*, 178–190.

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G., & Smolkowski, K. (2014). Regression models for the analysis of single case designs. *Neuropsychological Rehabilitation, 24*, 554–571.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and

n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*, 619–638.

Tukey, J. W. (1977). *Exploratory data analysis*. London, UK: Addison-Wesley.

Van Den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1–10.

Van Den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention, 2*, 142–151.

*van der Meer, L., Didden, R., Sutherland, D., O'Reilly, M. F., Lancioni, G. E., & Sigafoos, J. (2012). Comparing three augmentative and alternative communication modes for children with developmental disabilities. *Journal of Developmental and Physical Disabilities, 24*, 451–468.

*van der Meer, L., Didden, R., Sutherland, D., O'Reilly, M., Lancioni, G., & Sigafoos, J. (2013). Erratum to: Comparing three augmentative and alternative communication modes for children with developmental disabilities. *Journal of Developmental and Physical Disabilities, 25*, 271–272.

*van der Meer, L., Kagohara, D., Achmadi, D., O'Reilly, M. F., Lancioni, G. E., Sutherland, D., & Sigafoos, J. (2012). Speech-generating devices versus manual signing for children with developmental disabilities. *Research in Developmental Disabilities, 33*, 1658–1669.

*van der Meer, L., Kagohara, D., Roche, L., Sutherland, D., Balandin, S., Green, V. A., ... & Sigafoos, J. (2013). Teaching multi-step requesting and social communication to two children with autism spectrum disorders with three AAC options. *Augmentative and Alternative Communication, 29*, 222–234.

*van der Meer, L., Sutherland, D., O'Reilly, M. F., Lancioni, G. E., & Sigafoos, J. (2012). A further comparison of manual signing, picture exchange, and speech-generating devices as communication modes for children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 6*, 1247–1257.

*Van Laarhoven, T., Kraus, E., Karpman, K., Nizzi, R., & Valentino, J. (2010). A comparison of picture and video prompts to teach daily living skills to individuals with autism. *Focus on Autism and Other Developmental Disabilities, 25*, 195–208.

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development, 93*, 403–411.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., Nikles, J., Zucker, D. R., Kravitz, R., Guyatt, G., Altman, D. G., Moher, D., and the CENT group (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *British Medical Journal,* May 14; 350:h1738. doi: 10.1136/bmj.h1738.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. doi: 10.1177/1745691612463078

Westfall, P. H, & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for* p-*value adjustment.* New York, NY: Wiley.

Wolery, M. (2013). A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education, 43,* 39–43.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18–29.

Wolery, M., Gast, D. L., & Hammond, D. (2010). Comparative intervention designs. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 329–381). London, UK: Routledge.

*Yakubova, G., & Bouck, E. C. (2014). Not all created equally: Exploring calculator use by students with mild intellectual disability. *Education and Training in Autism and Developmental Disabilities, 49*, 111–126.

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage Publications.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173–181.

Running head: ATD DATA ANALYSIS

Figure captions:

Figure 1. Examples of real data gathered via alternating treatments designs.

Figure 2. Application of ADISO (average difference between successive observations) to the data gathered by Eilers and Hayes (2015) on the intervals of problem behavior by Jacob, comparing a control condition including exposure and a condition including both exposure and a cognitive defusion exercise: (A) performing comparisons only in one direction (all AB or all BA); (B) a user-defined set of comparisons.

Figure 3. Application of ALIV (average difference between linearly interpolated values) to the data gathered by Eilers and Hayes (2015) on the intervals of problem behavior by Jacob, comparing a control condition including exposure and a condition including both exposure and a cognitive defusion exercise. The arrows show the values that are actually being compared after linear interpolation (pointing up is deterioration, pointing down is improvement); the portion of the measurement occasions for which the comparisons are performed are marked by the vertical lines.